(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)公表番号
特表2023-541298
(P2023-541298A)
(43)公表日　令和5年9月29日(2023.9.29)

(51)国際特許分類		FI			テーマコード（参考）
G06F	*9/50 (2006.01)*	G06F	9/50	150D	5B175
G06F	*16/182 (2019.01)*	G06F	16/182		
G06F	*16/28 (2019.01)*	G06F	16/28		
G06F	*9/46 (2006.01)*	G06F	9/50	150A	
		G06F	9/46	430	

審査請求　有　予備審査請求　未請求　（全63頁）

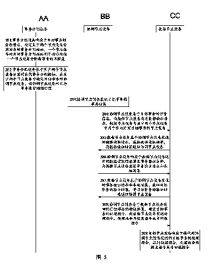
(21)出願番号　特願2023-517375(P2023-517375)
(86)(22)出願日　令和3年10月26日(2021.10.26)
(85)翻訳文提出日　令和5年3月15日(2023.3.15)
(86)国際出願番号　PCT/CN2021/126408
(87)国際公開番号　WO2022/111188
(87)国際公開日　令和4年6月2日(2022.6.2)
(31)優先権主張番号　202011362629.2
(32)優先日　令和2年11月27日(2020.11.27)
(33)優先権主張国・地域又は機関
中国(CN)
(81)指定国・地域　AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,
最終頁に続く

(71)出願人　517392436
▲騰▼▲訊▼科技（深▲セン▼）有限公司
中華人民共和国５１８０５７▲広▼▲東▼省深▲セン▼市南山区高新区科技中一路 ▲騰▼▲訊▼大厦３５▲層▼
(74)代理人　100110364
弁理士　実広 信哉
(74)代理人　100150197
弁理士　松尾 直樹
(72)発明者　李 ▲海▼翔
中華人民共和国５１８０５７▲広▼▲東▼省深▲セン▼市南山区高新区科技中一路 ▲騰▼▲訊▼大厦３５▲層▼

Fターム（参考）　5B175　AA01 EA03

(54)【発明の名称】　トランザクション処理方法、システム、装置、機器、及びプログラム

(57)【要約】
トランザクション処理方法、システム、装置、機器、記憶媒体及びプログラム製品は、データベースの技術分野に属する。方法は、ターゲットトランザクションの割り当て要求に応答して、少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定するステップ（２０１）と、少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づいて、少なくとも２つのノード機器のうちターゲットトランザクションの協調ノード機器を決定し、協調ノード機器によってターゲットトランザクションを協調処理するステップ（２０２）と、を含む。このような方式に基づいて、個々の協調ノード機器はいずれも分散化された機器としてトランザクションを協調処理することを可能にすることにより、トランザクションをノード間で処理することを可能にし、トランザクションの処理効率を向上させることに有利であり、トランザクション処理の信頼性が比較的高く、データベースシステムのシステム性能を高めることに有利である。



201 In response to an allocation request of a target transaction, a transaction allocation device determines transaction allocation indicators respectively corresponding to at least two node devices, the transaction allocation indicator corresponding to one node device being used for indicating a matching degree of allocating a new transaction to the node device
202 On the basis of the transaction allocation indicators respectively corresponding to the at least two node devices, the transaction allocation device determines a coordinator node device of the target transaction from the at least two node devices, so that the coordinator node device performs coordination processing on the target transaction
203 The coordinator node device obtains transaction information of the target transaction
204 On the basis of the transaction information of the target transaction, the coordinator node device sends a data reading request to the data node device, the data node device being a node device in the at least two node devices for participating in processing of the target transaction
205 On the basis of the data reading request sent by the coordinator node device, the data node device obtains a data reading result, and returns the data reading result to the coordinator node device
206 In response to the data reading result returned by the data node device satisfying a transaction verification condition, the coordinator node device sends a transaction verification request and a local write set to the data node device
207 On the basis of the transaction verification request and the local write set sent by the coordinator node device, the data node device obtains a verification result of the target transaction, and returns the verification result of the target transaction to the coordinator node device
208 On the basis of the verification result of the target transaction returned by the data node device, the coordinator node device determines a processing instruction of the target transaction, and sends the processing instruction to the data node device, the processing instruction being a submission instruction or a rollback instruction
209 In response to receiving the processing instruction of the target transaction from the coordinator node device, the data node device executes the processing instruction, the processing instruction being the submission instruction or the rollback instruction
AA Transaction allocation device
BB Coordinator node device
CC Data node device

10

20

【特許請求の範囲】
【請求項１】
トランザクション処理方法であって、前記方法はトランザクション割り当て機器に応用され、前記トランザクション割り当て機器は分散型データベースシステム中にあり、前記分散型データベースシステムにおいて同一の記憶システムを共有する少なくとも２つのノード機器がさらに含まれ、前記方法は、
ターゲットトランザクションの割り当て要求に応答して、前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定するステップであって、１つのノード機器に対応するトランザクション割り当て指標は前記１つのノード機器に新たなトランザクションを割り当てるマッチング度を指示することに用いられる、ステップと、
前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づいて、前記少なくとも２つのノード機器のうち前記ターゲットトランザクションの協調ノード機器を決定し、前記協調ノード機器によって前記ターゲットトランザクションを協調処理するステップと、を含む、トランザクション処理方法。
【請求項２】
前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定する前記ステップは、
トランザクション割り当てモードを決定するステップであって、前記トランザクション割り当てモードはトランザクションのビジー度に基づく割り当て、機器のビジー度に基づく割り当て、及びハイブリッドのビジー度に基づく割り当てのうちのいずれか１つを含む、ステップと、
前記トランザクション割り当てモードによって指示された決定方式に基づき、前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定するステップと、を含む、請求項１に記載の方法。
【請求項３】
前記トランザクション割り当てモードはハイブリッドのビジー度に基づく割り当てを含み、前記トランザクション割り当てモードによって指示された決定方式に基づき、前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定する前記ステップは、
第１ノード機器のトランザクション処理数、前記第１ノード機器の機器リソース使用率、トランザクション処理数の重み、機器リソース使用率の重み、及び重み調節パラメータに基づいて、前記第１ノード機器に対応するトランザクション割り当て指標を決定するステップであって、前記第１ノード機器は前記少なくとも２つのノード機器のうちのいずれか１つのノード機器である、ステップを含む、請求項２に記載の方法。
【請求項４】
前記少なくとも２つのノード機器のうち前記ターゲットトランザクションの協調ノード機器を決定する前記ステップの後に、前記方法は、
前記協調ノード機器の機器識別情報を前記割り当て要求を発した端末に送信するステップであって、前記端末は前記協調ノード機器の機器識別情報に基づき、前記ターゲットトランザクションのトランザクション情報を前記協調ノード機器に送信し、前記協調ノード機器によって前記トランザクション情報に基づいて前記ターゲットトランザクションを協調処理することに用いられる、ステップをさらに含む、請求項１～３のいずれか一項に記載の方法。
【請求項５】
前記分散型データベースシステムはキー値型のデータ記憶フォーマットとセグメントページ型のデータ記憶フォーマットをサポートする、請求項１～３のいずれか一項に記載の方法。
【請求項６】
トランザクション処理方法であって、前記方法は協調ノード機器に応用され、前記協調

ノード機器は同一の記憶システムを共有する少なくとも２つのノード機器のうちターゲットトランザクションを協調処理することに用いられるノード機器であり、前記協調ノード機器は前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づき決定され、前記方法は、

前記ターゲットトランザクションのトランザクション情報を取得するステップと、

前記ターゲットトランザクションのトランザクション情報に基づいて、データノード機器にデータ読み取り要求を送信するステップであって、前記データノード機器は前記少なくとも２つのノード機器のうち前記ターゲットトランザクションの処理に関与することに用いられるノード機器である、ステップと、

前記データノード機器から返信されたデータ読み取り結果がトランザクション認証条件を満たすことに応答して、前記データノード機器にトランザクション認証要求、及びローカル書き込みセットを送信するステップと、

前記データノード機器から返信された前記ターゲットトランザクションの認証結果に基づいて、前記ターゲットトランザクションの処理命令を決定し、前記データノード機器に前記処理命令を送信するステップであって、前記処理命令はコミット命令、又はアボート命令であり、前記データノード機器は前記処理命令を実行することに用いられる、ステップと、を含む、トランザクション処理方法。

【請求項７】

前記データ読み取り結果は第２論理ライフサイクルを携えており、前記第２論理ライフサイクルは前記データノード機器によって前記データ読み取り要求が携えている前記ターゲットトランザクションの第１論理ライフサイクルに基づき決定され、前記第１論理ライフサイクルはタイムスタンプ下限、及びタイムスタンプ上限で構成され、

前記データノード機器から返信されたデータ読み取り結果がトランザクション認証条件を満たすことに応答して、前記データノード機器にトランザクション認証要求を送信する前記ステップは、

前記第１論理ライフサイクルのタイムスタンプ下限と前記第２論理ライフサイクルのタイムスタンプ下限における最大値を前記ターゲットトランザクションの第３論理ライフサイクルのタイムスタンプ下限として用い、前記第１論理ライフサイクルのタイムスタンプ上限と前記第２論理ライフサイクルのタイムスタンプ上限における最小値を前記ターゲットトランザクションの第３論理ライフサイクルのタイムスタンプ上限として用いるステップと、

前記第３論理ライフサイクルが有効であることに応答して、前記データノード機器に前記第３論理ライフサイクルを携えているトランザクション認証要求を送信するステップであって、前記第３論理ライフサイクルが有効であることは、前記第３論理ライフサイクルのタイムスタンプ下限が前記第３論理ライフサイクルのタイムスタンプ上限よりも小さいことを指示することに用いられる、ステップと、を含む、請求項６に記載の方法。

【請求項８】

前記データノード機器の数は少なくとも２つであり、前記データノード機器から返信された前記ターゲットトランザクションの認証結果に基づいて、前記ターゲットトランザクションの処理命令を決定する前記ステップは、

前記少なくとも２つのデータノード機器から返信された少なくとも２つの認証結果において認証に合格しなかったことを指示することに用いられる認証結果が存在することに応答して、前記アボート命令を前記ターゲットトランザクションの処理命令として用いるステップと、

前記少なくとも２つのデータノード機器から返信された少なくとも２つの認証結果はいずれも認証に合格したことを指示することに応答して、前記少なくとも２つの認証結果が携えている論理ライフサイクルの共通部分をターゲット論理ライフサイクルとして用いるステップと、

前記ターゲット論理ライフサイクルが有効であることに応答して、前記コミット命令を前記ターゲットトランザクションの処理命令として用い、前記ターゲット論理ライフサイ

クルが無効であることに応答して、前記アボート命令を前記ターゲットトランザクションの処理命令として用いるステップと、を含む、請求項６又は７のいずれか一項に記載の方法。

【請求項９】

トランザクション処理方法であって、前記方法はデータノード機器に応用され、前記データノード機器は同一の記憶システムを共有する少なくとも２つのノード機器のうちターゲットトランザクションの処理に関与することに用いられるノード機器であり、前記方法は、

協調ノード機器から送信されたデータ読み取り要求に基づいて、データ読み取り結果を取得し、前記データ読み取り結果を前記協調ノード機器に返信するステップであって、前記協調ノード機器は前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づき決定される、ステップと、

前記協調ノード機器から送信されたトランザクション認証要求、及びローカル書き込みセットに基づいて、前記ターゲットトランザクションの認証結果を取得し、前記ターゲットトランザクションの認証結果を前記協調ノード機器に返信するステップと、

前記協調ノード機器から送信された前記ターゲットトランザクションの処理命令を受信したことに応答して、前記処理命令を実行するステップであって、前記処理命令はコミット命令、又はアボート命令である、ステップと、を含む、トランザクション処理方法。

【請求項１０】

前記データ読み取り要求は前記ターゲットトランザクションの第１論理ライフサイクルを携えており、前記第１論理ライフサイクルはタイムスタンプ下限、及びタイムスタンプ上限で構成され、協調ノード機器から送信されたデータ読み取り要求に基づいて、データ読み取り結果を取得する前記ステップは、

前記第１論理ライフサイクルに基づいて、前記データ読み取り要求によって指示された読み取り対象のデータ項目の可視バージョンデータを決定するステップと、

前記可視バージョンデータの作成タイムスタンプ、及び前記第１論理ライフサイクルに基づいて、前記ターゲットトランザクションの第２論理ライフサイクルを決定するステップと、

前記第２論理ライフサイクル、及び前記可視バージョンデータが運ばれた結果を前記データ読み取り結果として用いるステップと、を含む、請求項９に記載の方法。

【請求項１１】

前記トランザクション認証要求は前記ターゲットトランザクションの第３論理ライフサイクルを携えており、前記第３論理ライフサイクルは前記協調ノード機器によって前記第１論理ライフサイクル、及び前記第２論理ライフサイクルに基づいて決定された有効論理ライフサイクルであり、前記協調ノード機器から送信されたトランザクション認証要求、及びローカル書き込みセットに基づいて、前記ターゲットトランザクションの認証結果を取得する前記ステップは、

前記第３論理ライフサイクルのタイムスタンプ下限と前記第２論理ライフサイクルのタイムスタンプ下限における最大値を前記ターゲットトランザクションの第４論理ライフサイクルのタイムスタンプ下限として用い、前記第３論理ライフサイクルのタイムスタンプ上限と前記第２論理ライフサイクルのタイムスタンプ上限における最小値を前記ターゲットトランザクションの第４論理ライフサイクルのタイムスタンプ上限として用いるステップと、

前記第４論理ライフサイクルが有効であることに応答して、前記ローカル書き込みセットに対応する各書き込み対象のデータ項目の読み取りトランザクション関連情報及び前記第４論理ライフサイクルに基づいて、前記ターゲットトランザクションの第５論理ライフサイクルを決定するステップと、

前記第５論理ライフサイクルが有効であることに応答して、認証に合格したことを指示することに用いられる認証結果を前記ターゲットトランザクションの認証結果として用い、前記第５論理ライフサイクルが無効であることに応答して、認証に合格しなかったこと

を指示することに用いられる認証結果を前記ターゲットトランザクションの認証結果として用いるステップと、を含む、請求項１０に記載の方法。

【請求項１２】

１つの書き込み対象のデータ項目の読み取りトランザクション関連情報は前記１つの書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプを含み、前記１つの書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプは前記１つの書き込み対象のデータ項目を読み取った各読み取りトランザクションの論理コミットタイムスタンプにおける最大値を指示することに用いられ、前記ローカル書き込みセットに対応する各書き込み対象のデータ項目の読み取りトランザクション関連情報及び前記第４論理ライフサイクルに基づいて、前記ターゲットトランザクションの第５論理ライフサイクルを決定する前記ステップは、

前記各書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプ、及び前記第４論理ライフサイクルに基づいて、前記ターゲットトランザクションの第５論理ライフサイクルを決定するステップであって、前記第５論理ライフサイクルのタイムスタンプ下限は前記各書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプにおける最大値よりも大きい、ステップを含む、請求項１１に記載の方法。

【請求項１３】

１つの書き込み対象のデータ項目の読み取りトランザクション関連情報は前記１つの書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプを含み、前記ターゲット読み取りトランザクションはローカル認証に合格するか、又はコミット段階にある読み取りトランザクションであり、前記ターゲット読み取りトランザクションの終了タイムスタンプは前記ターゲット読み取りトランザクションの論理ライフサイクルのタイムスタンプ上限であり、前記ローカル書き込みセットに対応する各書き込み対象のデータ項目の読み取りトランザクション関連情報及び前記第４論理ライフサイクルに基づいて、前記ターゲットトランザクションの第５論理ライフサイクルを決定する前記ステップは、

前記各書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプ、及び前記第４論理ライフサイクルに基づいて、前記ターゲットトランザクションの第５論理ライフサイクルを決定するステップであって、前記第５論理ライフサイクルのタイムスタンプ下限は前記各書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプにおける最大値よりも大きい、ステップを含む、請求項１１に記載の方法。

【請求項１４】

トランザクション処理システムであって、前記トランザクション処理システムは協調ノード機器と、データノード機器とを含み、前記協調ノード機器は同一の記憶システムを共有する少なくとも２つのノード機器のうちターゲットトランザクションを協調処理することに用いられるノード機器であり、前記協調ノード機器は前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づき決定され、前記データノード機器は前記少なくとも２つのノード機器のうち前記ターゲットトランザクションの処理に関与することに用いられるノード機器であり、

前記協調ノード機器は、前記ターゲットトランザクションのトランザクション情報を取得することと、前記ターゲットトランザクションのトランザクション情報に基づいて、前記データノード機器にデータ読み取り要求を送信することと、に用いられ、

前記データノード機器は、前記協調ノード機器から送信された前記データ読み取り要求に基づいて、データ読み取り結果を取得し、前記データ読み取り結果を前記協調ノード機器に返信することに用いられ、

前記協調ノード機器はさらに、前記データノード機器から返信された前記データ読み取り結果がトランザクション認証条件を満たすことに応答して、前記データノード機器にトランザクション認証要求、及びローカル書き込みセットを送信することに用いられ、

前記データノード機器はさらに、前記協調ノード機器から送信された前記トランザクシ

ョン認証要求、及び前記ローカル書き込みセットに基づいて、前記ターゲットトランザクションの認証結果を取得し、前記ターゲットトランザクションの認証結果を前記協調ノード機器に返信することに用いられ、
前記協調ノード機器はさらに、前記データノード機器から返信された前記ターゲットトランザクションの認証結果に基づいて、前記ターゲットトランザクションの処理命令を決定し、前記データノード機器に前記処理命令を送信することに用いられ、前記処理命令はコミット命令、又はアボート命令であり、
前記データノード機器はさらに、前記協調ノード機器から送信された前記ターゲットトランザクションの処理命令を受信したことに応答して、前記処理命令を実行することに用いられる、トランザクション処理システム。

【請求項１５】
トランザクション処理装置であって、前記装置は、
ターゲットトランザクションの割り当て要求に応答して、同一の記憶システムを共有する少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定することに用いられる第１決定ユニットであって、１つのノード機器に対応するトランザクション割り当て指標は前記１つのノード機器に新たなトランザクションを割り当てるマッチング度を指示することに用いられる、第１決定ユニットと、
前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づいて、前記少なくとも２つのノード機器のうち前記ターゲットトランザクションの協調ノード機器を決定し、前記協調ノード機器によって前記ターゲットトランザクションを協調処理することに用いられる第２決定ユニットと、を含む、トランザクション処理装置。

【請求項１６】
トランザクション処理装置であって、前記装置は、
ターゲットトランザクションのトランザクション情報を取得することに用いられる取得ユニットと、
前記ターゲットトランザクションのトランザクション情報に基づいて、データノード機器にデータ読み取り要求を送信することに用いられる第１送信ユニットであって、前記データノード機器は同一の記憶システムを共有する少なくとも２つのノード機器のうち前記ターゲットトランザクションの処理に関与することに用いられるノード機器である、第１送信ユニットと、
前記データノード機器から返信されたデータ読み取り結果がトランザクション認証条件を満たすことに応答して、前記データノード機器にトランザクション認証要求及び、ローカル書き込みセットを送信することに用いられる第２送信ユニットと、
前記データノード機器から返信された前記ターゲットトランザクションの認証結果に基づいて、前記ターゲットトランザクションの処理命令を決定することに用いられる決定ユニットと、
前記データノード機器に前記処理命令を送信することに用いられる第３送信ユニットであって、前記処理命令はコミット命令、又はアボート命令であり、前記データノード機器は前記処理命令を実行することに用いられる、第３送信ユニットと、を含む、トランザクション処理装置。

【請求項１７】
トランザクション処理装置であって、前記装置は、
協調ノード機器から送信されたデータ読み取り要求に基づいて、データ読み取り結果を取得することに用いられる第１取得ユニットであって、前記協調ノード機器は同一の記憶システムを共有する少なくとも２つのノード機器のうちターゲットトランザクションを協調処理することに用いられるノード機器であり、前記協調ノード機器は前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づき決定される、第１取得ユニットと、
前記データ読み取り結果を前記協調ノード機器に返信することに用いられる返信ユニッ

トと、
前記協調ノード機器から送信されたトランザクション認証要求、及びローカル書き込みセットに基づいて、前記ターゲットトランザクションの認証結果を取得することに用いられる第２取得ユニットと、
前記返信ユニットはさらに、前記ターゲットトランザクションの認証結果を前記協調ノード機器に返信することに用いられ、
前記協調ノード機器から送信された前記ターゲットトランザクションの処理命令を受信したことに応答して、前記処理命令を実行することに用いられる実行ユニットであって、前記処理命令はコミット命令、又はアボート命令である、実行ユニットと、を含むトランザクション処理装置。
【請求項１８】
コンピュータ機器であって、前記コンピュータ機器はプロセッサと、メモリとを含み、前記メモリにおいて少なくとも１つのコンピュータプログラムが記憶され、前記少なくとも１つのコンピュータプログラムは前記プロセッサによってロードされて実行され、前記コンピュータ機器に、請求項１～５のいずれか一項に記載のトランザクション処理方法、又は請求項６～８のいずれか一項に記載のトランザクション処理方法、又は請求項９～１３のいずれか一項に記載のトランザクション処理方法を実現させる、コンピュータ機器。
【請求項１９】
非一時的コンピュータ可読記憶媒体であって、前記非一時的コンピュータ可読記憶媒体において少なくとも１つのコンピュータプログラムが記憶され、前記少なくとも１つのコンピュータプログラムはプロセッサによってロードされて実行され、コンピュータに、請求項１～５のいずれか一項に記載のトランザクション処理方法、又は請求項６～８のいずれか一項に記載のトランザクション処理方法、又は請求項９～１３のいずれか一項に記載のトランザクション処理方法を実現させる、非一時的コンピュータ可読記憶媒体。
【請求項２０】
コンピュータプログラム製品であって、前記コンピュータプログラム製品はコンピュータ命令を含み、前記コンピュータ命令はコンピュータ可読記憶媒体に記憶され、コンピュータ機器のプロセッサは前記コンピュータ可読記憶媒体から前記コンピュータ命令を読み取り、前記プロセッサが前記コンピュータ命令を実行することにより、前記コンピュータ機器が請求項１～５のいずれか一項に記載のトランザクション処理方法、請求項６～８のいずれか一項に記載のトランザクション処理方法、又は請求項９～１３のいずれか一項に記載のトランザクション処理方法を実行する、コンピュータプログラム製品。
【発明の詳細な説明】
【技術分野】
【０００１】
本願の実施例はデータベースの技術分野に関し、特にトランザクション処理方法、システム、装置、機器、記憶媒体及びプログラム製品に関する。
【０００２】
本願は、２０２０年１１月２７日に提出された、出願番号が第２０２０１１３６２６２９．２号、発明の名称が「トランザクション処理方法、機器及びコンピュータ可読記憶媒体」の中国特許出願の優先権を主張し、その全内容は引用により本願に組み込まれている。
【背景技術】
【０００３】
データベース技術の発展に伴って、ビッグデータ、クラウドコンピューティングなどのビジネスシナリオに適応することを可能にするために、分散型データベースシステムが徐々に普及している。複数種の分散型データベースシステムにおいて、共有記憶（ｓｈａｒｅ－ｄｉｓｋ）アーキテクチャに基づく分散型データベースシステムは主流のシステムとなっている。
【０００４】

現在、ｓｈａｒｅ－ｄｉｓｋアーキテクチャに基づく分散型データベースシステムにおいて、データ項目の分布状況に基づきトランザクションを割り当てて、あるデータ項目に関するトランザクションを該データ項目にサービスされる固定ノード機器に割り当てて独立して処理する。これに基づいて、トランザクションの処理効率が大幅に制限されてしまい、トランザクション処理の信頼性が比較的悪くなる。

【発明の概要】

【発明が解決しようとする課題】

【０００５】

本願の実施例はトランザクション処理方法、システム、装置、機器、記憶媒体及びプログラム製品を提供し、トランザクションの処理効率を向上させることに用いることができる。

【課題を解決するための手段】

【０００６】

一態様では、本願の実施例はトランザクション処理方法を提供し、前記方法はトランザクション割り当て機器に応用され、前記トランザクション割り当て機器は分散型データベースシステム中にあり、前記分散型データベースシステムにおいて同一の記憶システムを共有する少なくとも２つのノード機器がさらに含まれ、前記方法は、

ターゲットトランザクションの割り当て要求に応答して、前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定するステップであって、１つのノード機器に対応するトランザクション割り当て指標は前記１つのノード機器に新たなトランザクションを割り当てるマッチング度を指示することに用いられる、ステップと、

前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づいて、前記少なくとも２つのノード機器のうち前記ターゲットトランザクションの協調ノード機器を決定し、前記協調ノード機器によって前記ターゲットトランザクションを協調処理するステップと、を含む。

【０００７】

トランザクション処理方法をさらに提供し、前記方法は協調ノード機器に応用され、前記協調ノード機器は同一の記憶システムを共有する少なくとも２つのノード機器のうちターゲットトランザクションを協調処理することに用いられるノード機器であり、前記協調ノード機器は前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づき決定され、前記方法は、

前記ターゲットトランザクションのトランザクション情報を取得するステップと、

前記ターゲットトランザクションのトランザクション情報に基づいて、データノード機器にデータ読み取り要求を送信するステップであって、前記データノード機器は前記少なくとも２つのノード機器のうち前記ターゲットトランザクションの処理に関与することに用いられるノード機器である、ステップと、

前記データノード機器から返信されたデータ読み取り結果がトランザクション認証条件を満たすことに応答して、前記データノード機器にトランザクション認証要求、及びローカル書き込みセットを送信するステップと、

前記データノード機器から返信された前記ターゲットトランザクションの認証結果に基づいて、前記ターゲットトランザクションの処理命令を決定し、前記データノード機器に前記処理命令を送信するステップであって、前記処理命令はコミット命令、又はアボート命令であり、前記データノード機器は前記処理命令を実行することに用いられる、ステップと、を含む。

【０００８】

トランザクション処理方法をさらに提供し、前記方法はデータノード機器に応用され、前記データノード機器は同一の記憶システムを共有する少なくとも２つのノード機器のうちターゲットトランザクションの処理に関与することに用いられるノード機器であり、前記方法は、

協調ノード機器から送信されたデータ読み取り要求に基づいて、データ読み取り結果を取得し、前記データ読み取り結果を前記協調ノード機器に返信するステップであって、前記協調ノード機器は前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づき決定される、ステップと、
前記協調ノード機器から送信されたトランザクション認証要求、及びローカル書き込みセットに基づいて、前記ターゲットトランザクションの認証結果を取得し、前記ターゲットトランザクションの認証結果を前記協調ノード機器に返信するステップと、
前記協調ノード機器から送信された前記ターゲットトランザクションの処理命令を受信したことに応答して、前記処理命令を実行するステップであって、前記処理命令はコミット命令、又はアボート命令である、ステップと、を含む。

【０００９】

別の態様では、トランザクション処理システムを提供し、前記トランザクション処理システムは協調ノード機器と、データノード機器とを含み、前記協調ノード機器は同一の記憶システムを共有する少なくとも２つのノード機器のうちターゲットトランザクションを協調処理することに用いられるノード機器であり、前記協調ノード機器は前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づき決定され、前記データノード機器は前記少なくとも２つのノード機器のうち前記ターゲットトランザクションの処理に関与することに用いられるノード機器であり、
前記協調ノード機器は、前記ターゲットトランザクションのトランザクション情報を取得することと、前記ターゲットトランザクションのトランザクション情報に基づいて、前記データノード機器にデータ読み取り要求を送信することと、に用いられ、
前記データノード機器は、前記協調ノード機器から送信された前記データ読み取り要求に基づいて、データ読み取り結果を取得し、前記データ読み取り結果を前記協調ノード機器に返信することに用いられ、
前記協調ノード機器はさらに、前記データノード機器から返信された前記データ読み取り結果がトランザクション認証条件を満たすことに応答して、前記データノード機器にトランザクション認証要求、及びローカル書き込みセットを送信することに用いられ、
前記データノード機器はさらに、前記協調ノード機器から送信された前記トランザクション認証要求、及び前記ローカル書き込みセットに基づいて、前記ターゲットトランザクションの認証結果を取得し、前記ターゲットトランザクションの認証結果を前記協調ノード機器に返信することに用いられ、
前記協調ノード機器はさらに、前記データノード機器から返信された前記ターゲットトランザクションの認証結果に基づいて、前記ターゲットトランザクションの処理命令を決定し、前記データノード機器に前記処理命令を送信することに用いられ、前記処理命令はコミット命令、又はアボート命令であり、
前記データノード機器はさらに、前記協調ノード機器から送信された前記ターゲットトランザクションの処理命令を受信したことに応答して、前記処理命令を実行することに用いられる。

【００１０】

別の態様では、トランザクション処理装置を提供し、前記装置は、
ターゲットトランザクションの割り当て要求に応答して、同一の記憶システムを共有する少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定することに用いられる第１決定ユニットであって、１つのノード機器に対応するトランザクション割り当て指標は前記１つのノード機器に新たなトランザクションを割り当てるマッチング度を指示することに用いられる、第１決定ユニットと、
前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づいて、前記少なくとも２つのノード機器のうち前記ターゲットトランザクションの協調ノード機器を決定し、前記協調ノード機器によって前記ターゲットトランザクションを協調処理することに用いられる第２決定ユニットと、を含む。

【００１１】

トランザクション処理装置をさらに提供し、前記装置は、
ターゲットトランザクションのトランザクション情報を取得することに用いられる取得ユニットと、
前記ターゲットトランザクションのトランザクション情報に基づいて、データノード機器にデータ読み取り要求を送信することに用いられる第１送信ユニットであって、前記データノード機器は同一の記憶システムを共有する少なくとも２つのノード機器のうち前記ターゲットトランザクションの処理に関与することに用いられるノード機器である、第１送信ユニットと、
前記データノード機器から返信されたデータ読み取り結果がトランザクション認証条件を満たすことに応答して、前記データノード機器にトランザクション認証要求、及びローカル書き込みセットを送信することに用いられる第２送信ユニットと、
前記データノード機器から返信された前記ターゲットトランザクションの認証結果に基づいて、前記ターゲットトランザクションの処理命令を決定することに用いられる決定ユニットと、
前記データノード機器に前記処理命令を送信することに用いられる第３送信ユニットであって、前記処理命令はコミット命令、又はアボート命令であり、前記データノード機器は前記処理命令を実行することに用いられる、第３送信ユニットと、を含む。

【００１２】

トランザクション処理装置をさらに提供し、前記装置は、
協調ノード機器から送信されたデータ読み取り要求に基づいて、データ読み取り結果を取得することに用いられる第１取得ユニットであって、前記協調ノード機器は同一の記憶システムを共有する少なくとも２つのノード機器のうちターゲットトランザクションを協調処理することに用いられるノード機器であり、前記協調ノード機器は前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づき決定される、第１取得ユニットと、
前記データ読み取り結果を前記協調ノード機器に返信することに用いられる返信ユニットと、
前記協調ノード機器から送信されたトランザクション認証要求、及びローカル書き込みセットに基づいて、前記ターゲットトランザクションの認証結果を取得することに用いられる第２取得ユニットと、
前記協調ノード機器から送信された前記ターゲットトランザクションの処理命令を受信したことに応答して、前記処理命令を実行することに用いられる実行ユニットであって、前記処理命令はコミット命令、又はアボート命令である、実行ユニットと、を含み、
前記返信ユニットはさらに、前記ターゲットトランザクションの認証結果を前記協調ノード機器に返信することに用いられる。

【００１３】

別の態様では、コンピュータ機器を提供し、前記コンピュータ機器はプロセッサと、メモリとを含み、前記メモリにおいて少なくとも１つのコンピュータプログラムが記憶され、前記少なくとも１つのコンピュータプログラムは前記プロセッサによってロードされて実行され、前記コンピュータ機器に前記のいずれか一項に記載のトランザクション処理方法を実現させる。

【００１４】

別の態様では、非一時的コンピュータ可読記憶媒体をさらに提供し、前記非一時的コンピュータ可読記憶媒体において少なくとも１つのコンピュータプログラムが記憶され、前記少なくとも１つのコンピュータプログラムはプロセッサによってロードされて実行され、コンピュータに前記のいずれか一項に記載のトランザクション処理方法を実現させる。

【００１５】

別の態様では、コンピュータプログラム製品、又はコンピュータプログラムをさらに提供し、前記コンピュータプログラム製品、又はコンピュータプログラムはコンピュータ命令を含み、前記コンピュータ命令はコンピュータ可読記憶媒体に記憶され、コンピュータ

機器のプロセッサは前記コンピュータ可読記憶媒体から前記コンピュータ命令を読み取り、前記プロセッサは前記コンピュータ命令を実行することにより、前記コンピュータ機器が前記のいずれか一項に記載のトランザクション処理方法を実行する。

【００１６】

本願の実施例において、各ノード機器のそれぞれに対応するトランザクション割り当て指標に基づきターゲットトランザクションを協調処理することに用いられる協調ノード機器を決定し、トランザクションの割り当て過程はトランザクションに関するデータ項目を考慮する必要がなく、データ項目の分布状況を考慮する必要もない。このような方式に基づいて、個々のノード機器はいずれも分散化された機器としてトランザクションを協調処理することを可能にすることにより、トランザクションをノード間で処理することを可能にし、トランザクションの処理効率を向上させることに有利となり、トランザクション処理の信頼性が比較的高くなり、データベースシステムのシステム性能を高めることに有利である。

【図面の簡単な説明】

【００１７】

【図１】本願の実施例が提供するトランザクション処理方法の実施環境の模式図である。

【図２】本願の実施例が提供するトランザクション処理方法のフローチャートである。

【図３】本願の実施例が提供するトランザクションログのフォーマットの模式図である。

【図４】本願の実施例が提供するトランザクションログのフォーマットの模式図である。

【図５】本願の実施例が提供するトランザクション処理装置の模式図である。

【図６】本願の実施例が提供するトランザクション処理装置の模式図である。

【図７】本願の実施例が提供するトランザクション処理装置の模式図である。

【図８】本願の実施例が提供するコンピュータ機器の構造模式図である。

【発明を実施するための形態】

【００１８】

本願の目的、技術的解決手段及び利点をより明確にするために、以下、図面を組み合わせて本願の実施形態をさらに詳細に記述する。

【００１９】

説明する必要があるように、本願の明細書及び特許請求の範囲における用語「第１」、「第２」などは、類似の対象を区別することに用いられるものであり、特定の順序又は前後順番を記述することに用いられるものではない。理解すべきであるように、ここで記述された本願の実施例がここで図示又は記述されたそれらのもの以外の順序で実施されることを可能にするように、このように使用されるデータは適切な場合に相互に交換することができる。以下の例示的な実施例において記述された実施形態は本願と一致するすべての実施形態を代表するものではない。逆に、これらは添付の特許請求の範囲において詳述された本願のいくつかの態様と一致する装置及び方法の例に過ぎない。

【００２０】

いくつかの実施例において、本願の実施例に関する分散型データベースシステムは、共有記憶（ｓｈａｒｅ－ｄｉｓｋ）アーキテクチャに基づく分散型データベースシステムであり、共有記憶アーキテクチャに基づく分散型データベースシステムにおいて少なくとも２つのノード機器が含まれ、該少なくとも２つのノード機器は自体のローカルの内部メモリエリアを有し、ネットワーク通信メカニズムによって同一の記憶システムに直接アクセスし、すなわち少なくとも２つのノード機器は同一の記憶システムを共有する。例えば、同一のＨＤＦＳ（Ｈａｄｏｏｐ Ｄｉｓｔｒｉｂｕｔｅｄ Ｆｉｌｅ Ｓｙｓｔｅｍ、分散型ファイルシステム）を共有する。少なくとも２つのノード機器によって共有された記憶システムにおいて複数のデータテーブルを記憶することができ、個々のデータテーブルは１つ又は複数のデータ項目を記憶することに用いることができる。

【００２１】

論理という角度から見ると、分散型データベースシステムにおけるノード機器は、協調ノード機器及びデータノード機器の２つの種類のロールに分割することができる。ここで

、協調ノード機器は主に生産、配信処理計画、及び分散型トランザクションの協調を担当し、データノード機器は主に協調ノード機器から送信された処理計画を受信し、相応なトランザクションを実行して協調ノード機器にトランザクションに関する関連データを返信することを担当する。

【００２２】

分散型データベースシステムにおいて、最小の操作実行ユニットはトランザクションであり、トランザクションが複数のデータノード機器上のデータ項目を操作する必要があるか否かに応じて、トランザクションは分散型トランザクション及びローカルトランザクションの２つの種類に分割することができる。この２つの種類の異なるトランザクションに対しては、それぞれ異なる実行プロセスを採用して、ネットワーク通信オーバーヘッドをできるだけ減少させ、トランザクション処理効率を高めることができる。ここで、分散型トランザクションは複数のデータノード機器間で読み書き操作を実行する必要があるトランザクションを表し、すなわちトランザクションは複数のデータノード機器上のデータ項目を操作する必要がある。たとえば、トランザクションＴはデータノード機器ＲＭ１、ＲＭ２、ＲＭ３上のデータ項目を操作する必要がある場合、該トランザクションＴは１つの分散型トランザクションである。ローカルトランザクションは単一のデータノード機器上のデータ項目のみを操作する必要があるトランザクションを表し、例えば、トランザクションＴはＲＭ１上のデータ項目のみを操作する必要があれば、該トランザクションＴは１つのローカルトランザクションである。

【００２３】

図１は本願の実施例が提供するトランザクション処理方法の実施環境の模式図である。図１に参照されるように、本願の実施例はｓｈａｒｅ－ｄｉｓｋフレームワークに基づく分散型データベースシステムに応用されてもよく、該分散型データベースシステムにおいてゲートウェイサーバ１０１と、トランザクション割り当て機器１０２と、分散型記憶クラスタ１０３と、グローバルタイムスタンプ生成クラスタ１０４とが含まれてもよい。分散型記憶クラスタ１０３においてｍ（ｍは２以上の整数である）個のノード機器が含まれ、該ｍ個のノード機器は同一の記憶システムを共有する。

【００２４】

ゲートウェイサーバ１０１は外部の読み書き要求を受信し、かつ読み書き要求に対応する読み書きトランザクションをトランザクション割り当て機器１０２、又は、分散型記憶クラスタ１０３に配信することに用いられる。たとえば、ユーザは端末上のアプリケーションクライアント端末にログインした後、アプリケーションクライアント端末をトリガーして読み書き要求を生成し、分散型データベースシステムによって提供されたＡＰＩ（Ａｐｐｌｉｃａｔｉｏｎ　Ｐｒｏｇｒａｍｍｉｎｇ　Ｉｎｔｅｒｆａｃｅ、アプリケーションプログラムプログラミングインタフェース）を呼び出して該読み書き要求に対応する読み書きトランザクションをゲートウェイサーバ１０１に送信する。

【００２５】

いくつかの実施例において、ゲートウェイサーバ１０１は分散型記憶クラスタ１０３におけるいずれか１つのノード機器と同一の物理マシン上に合併されてもよく、すなわち、あるノード機器をゲートウェイサーバ１０１として機能させる。

【００２６】

いくつかの実施例において、アプリケーションクライアント端末が所在する端末は分散型データベースシステムにおけるトランザクション割り当て機器１０２、及び、分散型記憶クラスタ１０３と通信接続を直接確立することを可能にする。このような場合に、分散型データベースシステムにおいてゲートウェイサーバ１０１が存在しなくてもよい。

【００２７】

トランザクション割り当て機器１０２は新たなトランザクションに適切なノード機器を協調ノード機器として割り当てることに用いられる。例示的な実施例において、トランザクション割り当て機器は分散型協調システム（たとえばＺｏｏＫｅｅｐｅｒ）にある。分散型協調システムはゲートウェイサーバ１０１、分散型記憶クラスタ１０３、及びグロー

バルタイムスタンプ生成クラスタ１０４のうちの少なくとも１つを管理することに用いられてもよい。選択的に、技術者は、端末上のスケジューラ（ｓｃｈｅｄｕｌｅｒ）を介して該分散型協調システムにアクセスし、それによりフロントエンドのスケジューラに基づいてバックエンドの分散型協調システムを制御して、各クラスタ又はサーバに対する管理を実現することができる。例えば、技術者は、スケジューラを介してＺｏｏＫｅｅｐｅｒを制御してあるノード機器を分散型記憶クラスタ１０３から削除することができ、すなわちあるノード機器を無効にすることができる。

【００２８】

分散型記憶クラスタ１０３はデータノード機器と、協調ノード機器とを含んでもよく、個々の協調ノード機器は少なくとも１つのデータノード機器に対応してもよい。データノード機器と協調ノード機器の分割は異なるトランザクションに対するものである。ある分散型トランザクションを例とすると、分散型トランザクションを発したノード機器は協調ノード機器と呼ばれてもよく、分散型トランザクションに関する他のノード機器はデータノード機器と呼ばれる。データノード機器又は協調ノード機器の数は１つ又は複数であってもよく、本願の実施例は分散型記憶クラスタ１０３におけるデータノード機器又は協調ノード機器の数を具体的に限定しない。

【００２９】

本願の実施例が提供する分散型データベースシステムにおいては、グローバルトランザクションマネージャがないため、該システムにおいてＸＡ（ｅＸｔｅｎｄｅｄ　Ａｒｃｈｉｔｅｃｔｕｒｅ、Ｘ／Ｏｐｅｎ組織の分散型トランザクション規範）／２ＰＣ（Ｔｗｏ－Ｐｈａｓｅ　Ｃｏｍｍｉｔ、２相コミット）技術を採用してノード間のトランザクション（分散型トランザクション）をサポートし、ノード間での書き込み操作時のデータの原子性及び一致性を確保することができる。このとき、協調ノード機器は２ＰＣアルゴリズムにおけるコーディネーターとして機能することに用いられ、該協調ノード機器と対応する各データノード機器は２ＰＣアルゴリズムにおける関与者として機能することに用いられる。

【００３０】

個々のデータノード機器、又は、協調ノード機器はスタンドアロン機器であってもよく、マスタースタンバイ構造（すなわちワンマスターマルチスタンバイクラスタ）を採用してもよい。図１に示すように、ノード機器（データノード機器又は協調ノード機器）がワンマスターデュアルスタンバイクラスタであることを例として例示すると、個々のノード機器においては１つのマスター、及び２つのスタンバイが含まれ、選択的に、個々のマスター又はスタンバイはいずれもエージェント（ａｇｅｎｔ）機器が対応して配置される。エージェント機器はマスター又はスタンバイと物理的に独立してもよく、もちろん、エージェント機器はさらにマスター又はスタンバイにおける１つのエイジェントモジュールとして用いられてもよい。ノード機器１を例とすると、ノード機器１は１つのマスターデータベース、及びエージェント機器（マスターｄａｔａｂａｓｅ＋ａｇｅｎｔ、マスターＤＢ＋ａｇｅｎｔと略称する）を含み、この他、２つのスタンバイデータベース及びエージェント機器（スタンバイｄａｔａｂａｓｅ＋ａｇｅｎｔ、スタンバイＤＢ＋ａｇｅｎｔと略称する）をさらに含む。マスターデータベースは上記したマスターであり、スタンバイデータベースは上記したスタンバイである。

【００３１】

グローバルタイムスタンプ生成クラスタ１０４は分散型トランザクションのグローバルコミットタイムスタンプ（Ｇｌｏｂａｌ　Ｔｉｍｅｓｔａｍｐ、Ｇｔｓ）を生成することに用いられ、該分散型トランザクションは複数のデータノード機器に関するトランザクションを指してもよく、例えば分散型読み取りトランザクションは複数のデータノード機器に記憶されたデータの読み取りに関してもよい。さらに例えば、分散型書き込みトランザクションは複数のデータノード機器におけるデータの書き込みに関してもよい。グローバルタイムスタンプ生成クラスタ１０４は、論理的に１つのシングルポイントとみなされてもよいが、いくつかの実施例において、１マスター３スレーブのアーキテクチャによって

より高い可用性を有するサービスを提供でき、クラスタの形式を採用して該グローバルコミットタイムスタンプの生成を実現することにより、シングルポイントの故障を防止し、それによりシングルポイントのボトルネック問題を避けることができる。

【００３２】

選択的に、グローバルコミットタイムスタンプは、分散型データベースシステムにおいてグローバルに一意的でかつ単調に逓増する１つのタイムスタンプ識別子であり、個々のトランザクションのグローバルコミット順序をマークする。これによりトランザクション間の実際の時間上の前後関係（トランザクションの全順序関係）を反映することに用いることを可能にし、グローバルコミットタイムスタンプは物理クロック、論理クロック又はハイブリッド物理クロックのうちの少なくとも１つを採用してもよく、本願の実施例はグローバルコミットタイムスタンプのタイプを具体的に限定しない。

【００３３】

いくつかの実施例において、該グローバルタイムスタンプ生成クラスタ１０４は物理的に独立してもよく、分散型協調システム（例えばＺｏｏＫｅｅｐｅｒ）と合併されてもよい。

【００３４】

上記図１は軽量なトランザクション処理を提供するアーキテクチャ図に過ぎず、ｓｈａｒｅ－ｄｉｓｋアーキテクチャに基づく分散型データベースシステムの１つの例示的な記述である。いくつかの実施例において、上記ゲートウェイサーバ１０１、トランザクション割り当て機器１０２、分散型記憶クラスタ１０３、及びグローバルタイムスタンプ生成クラスタ１０４で構成された分散型データベースシステムは、ユーザ端末にデータサービスを提供するサーバとみなされてもよい。該サーバは独立した物理サーバであってもよく、複数の物理サーバで構成されたサーバクラスタ又は分散型システムであってもよく、クラウドサービス、クラウドデータベース、クラウドコンピューティング、クラウド関数、クラウド記憶、ネットワークサービス、クラウド通信、ミドルウェアサービス、ドメイン名サービス、セキュリティサービス、ＣＤＮ（Ｃｏｎｔｅｎｔ Ｄｅｌｉｖｅｒｙ Ｎｅｔｗｏｒｋ、コンテンツ配信ネットワーク）、及びビッグデータ、並びに人工知能プラットフォームなどの基礎的なクラウドコンピューティングサービスを提供するクラウドサーバであってもよい。選択的に、上記ユーザ端末はスマートフォン、タブレットコンピュータ、ノートパソコン、デスクトップコンピュータ、スマートスピーカー、スマートウォッチなどであってもよいが、これらに限定されない。端末、及びサーバは有線又は無線通信方式によって直接又は間接的に接続されてもよく、本願はここで制限しない。

【００３５】

上記図１に示される実施環境に基づいて、本願の実施例はトランザクション処理方法を提供する。図２に示すように、本願の実施例が提供する方法は以下のステップ２０１～ステップ２０９を含む。

【００３６】

ステップ２０１において、トランザクション割り当て機器は、ターゲットトランザクションの割り当て要求に応答して、少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定し、１つのノード機器に対応するトランザクション割り当て指標は該１つのノード機器に新たなトランザクションを割り当てるマッチング度を指示することに用いられる。

【００３７】

トランザクション割り当て機器、及び少なくとも２つのノード機器はいずれも分散型データベースシステム中にあり、かつ少なくとも２つのノード機器は同一の記憶システムを共有する。本願の実施例は分散型データシステムの具体的な構造を限定せず、トランザクション割り当て機器と、同一の記憶システムを共有する少なくとも２つのノード機器とを含むものであればよい。

【００３８】

ターゲットトランザクションとは処理対象のトランザクションを指し、ターゲットトラ

ンザクションは分散型トランザクションであってもよく、ローカルトランザクションであってもよいが、本願の実施例はこれを限定しない。ターゲットトランザクションの割り当て要求はターゲットトランザクションに適切なノード機器を協調ノード機器として割り当てて、割り当てられた協調ノード機器によって該ターゲットトランザクションを協調処理するように指示することに用いられる。

【００３９】

ターゲットトランザクションの割り当て要求は端末によって発され、端末によって発されたターゲットトランザクションの割り当て要求は端末によってトランザクション割り当て機器に直接送信されるか、又はゲートウェイサーバによってトランザクション割り当て機器に転送されるが、本願の実施例はこれを限定しない。端末はユーザと対応するいずれかの電子機器であってもよく、スマートフォン、タブレットコンピュータ、ノートパソコン、デスクトップコンピュータ、スマートスピーカー、又はスマートウォッチのうちの少なくとも１つを含むがこれらに限定されず、本願の実施例は端末のタイプを具体的に限定しない。選択的に、端末上にアプリケーションクライアント端末がインストールされ、該アプリケーションクライアント端末はデータサービスを提供することを可能にするいずれかのクライアント端末であってもよく、例えば、該アプリケーションクライアント端末は支払いアプリケーションクライアント端末、フードデリバリーアプリケーションクライアント端末、タクシーアプリケーションクライアント端末又はソーシャルアプリケーションクライアント端末のうちの少なくとも１つであってもよく、本願の実施例はアプリケーションクライアント端末のタイプを具体的に限定しない。

【００４０】

少なくとも２つのノード機器とは、分散型データベースシステムにおいて分散化されたノード機器としてトランザクションを協調処理することを可能にするノード機器を指し、個々のノード機器はいずれも分散化されたアルゴリズムによって分散型トランザクションを協調処理することに用いることを可能にする。

【００４１】

トランザクション割り当て機器は、ターゲットトランザクションの割り当て要求を受信した後、ターゲットトランザクションに適切なノード機器を協調ノード機器として割り当てて、トランザクション処理の効率を確保する必要がある。ターゲットトランザクションに適切なノード機器を協調ノード機器として割り当てる過程において、トランザクション割り当て機器は先ず少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定する。１つのノード機器に対応するトランザクション割り当て指標は該１つのノード機器に新たなトランザクションを割り当てるマッチング度を指示することに用いられる。１つのノード機器に新たなトランザクションを割り当てるマッチング度が高いほど、該１つのノード機器に新たなトランザクションを割り当てるのにより適していることが説明される。

【００４２】

トランザクション割り当て指標は、トランザクションの観点から決定された、あるノード機器に新たなトランザクションを割り当てるのに適しているか否かを評価することに用いられる指標である。１つの可能な実現形態において、少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定する過程は以下のステップ２０１１とステップ２０１２とを含む。

【００４３】

ステップ２０１１：トランザクション割り当てモードを決定し、トランザクション割り当てモードはトランザクションのビジー度に基づく割り当て、機器のビジー度に基づく割り当て、及びハイブリッドのビジー度に基づく割り当てのうちのいずれか１つを含む。

【００４４】

トランザクション割り当てモードはノード機器に対応するトランザクション割り当て指標を決定する決定方式を指示することに用いられる。いくつかの実施例において、トランザクション割り当てモードは開発者によって設定されてトランザクション割り当て機器に

アップロードされる。説明する必要があるように、異なる時期に採用されたトランザクション割り当てモードは異なる可能性がある。ここでは、ステップ２０１１において決定されるのは、ターゲットトランザクションの割り当て要求を受信したときに採用すべきトランザクション割り当てモードである。

【００４５】

トランザクション割り当てモードはトランザクションのビジー度に基づく割り当て、機器のビジー度に基づく割り当て、及びハイブリッドのビジー度に基づく割り当てのうちのいずれか１つを含む。ここで、トランザクションのビジー度に基づく割り当てモードとは、ノード機器のトランザクション処理数を考慮する観点からトランザクション割り当て指標を決定することを指し、ノード機器のトランザクション処理数はノード機器のトランザクションのビジー度を反映することを可能にする。機器のビジー度に基づく割り当てモードとは、ノード機器の機器リソース使用率を考慮する観点からトランザクション割り当て指標を決定することを指し、ノード機器の機器リソース使用率はノード機器の機器のビジー度を反映することを可能にする。ハイブリッドのビジー度に基づく割り当てモードとは、ノード機器のトランザクション処理数、及びノード機器の機器リソース使用率を総合的に考慮する観点からトランザクション割り当て指標を決定することを指し、ノード機器のトランザクション処理数、及びノード機器の機器リソース使用率はノード機器のハイブリッドのビジー度を反映することを可能にする。

【００４６】

ステップ２０１２：トランザクション割り当てモードによって指示された決定方式に基づき、少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定する。

【００４７】

異なるトランザクション割り当てモードによって指示された決定方式は異なり、トランザクション割り当てモードを決定した後、トランザクション割り当てモードによって指示された決定方式に基づき、少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定する。次に、異なるトランザクション割り当てモード下で、少なくとも２つのノード機器のうちの第１ノード機器に対応するトランザクション割り当て指標を決定する方式をそれぞれ紹介する。ここで、第１ノード機器は少なくとも２つのノード機器のうちのいずれか１つのノード機器である。

【００４８】

いくつかの実施例において、トランザクション割り当てモードはトランザクションのビジー度に基づく割り当てである。このような場合に、トランザクション割り当てモードによって指示された決定方式に基づき、第１ノード機器に対応するトランザクション割り当て指標を決定する方式としては、該第１ノード機器のトランザクション処理数に基づいて、該第１ノード機器に対応するトランザクション割り当て指標を決定する。

【００４９】

第１ノード機器のトランザクション処理数とは、第１ノード機器が単位時間当たりに処理する必要があるトランザクションの数を指す。説明する必要があるように、ここでの処理する必要があるトランザクションとは、既に該第１ノード機器に割り当てられて処理されたトランザクションを指す。第１ノード機器が単位時間当たりに処理する必要があるトランザクションの数が多いほど、該第１ノード機器に新たなトランザクションを割り当てるのに適しないことが説明される。例示的な実施例において、第１ノード機器のトランザクション処理数は該第１ノード機器によってトランザクション割り当て機器にフィードバックされてもよく、トランザクション割り当て機器によってトランザクション割り当て状況に基づき自ら決定されてもよいが、本願の実施例はこれを限定しない。

【００５０】

本願の実施例はトランザクション割り当て指標の表現形式を限定せず、例示的には、トランザクション割り当て指標の表現形式はビジーレベル、又は数値である。

【００５１】

例示的には、トランザクション割り当て指標の表現形式がビジーレベルである場合、該第１ノード機器のトランザクション処理数に基づいて該第１ノード機器に対応するトランザクション割り当て指標を決定する方式としては、異なるビジーレベルに異なるトランザクション処理数の範囲を設定し、該第１ノード機器のトランザクション処理数があるトランザクション処理範囲に対応するビジーレベルを該第１ノード機器に対応するビジーレベルとして用いる。例示的には、ビジーレベルは「ビジー」、「部分ビジー」及び「アイドル」を含み、「ビジー」に対応するトランザクション処理数の範囲は［１０，＋∞）であり、「部分ビジー」に対応するトランザクション処理数の範囲は［３，１０）であり、「アイドル」に対応するトランザクション処理数の範囲は［０，３）である。第１ノード機器のトランザクション処理数が２であれば、「アイドル」を該第１ノード機器に対応するトランザクション割り当て指標として用いる。第１ノード機器に対応するトランザクション割り当て指標が「アイドル」に近くなるほど、該第１ノード機器に新たなトランザクションを割り当てるマッチング度が高くなることが説明される。

【００５２】

例示的には、トランザクション割り当て指標の表現形式が数値である場合、該第１ノード機器のトランザクション処理数に基づいて該第１ノード機器に対応するトランザクション割り当て指標を決定する方式としては、該第１ノード機器のトランザクション処理数を数値化処理し、数値化処理した後に得られた数値を該第１ノード機器に対応するトランザクション割り当て指標として用いる。トランザクション処理数を数値化処理する方式は経験に基づき設定されるか、又は応用シナリオに基づき柔軟に調整されるが、本願の実施例はこれを限定しない。例示的には、トランザクション処理数を数値化処理する方式としては、トランザクション処理数と参照重みとの積を計算する。このような方式では、トランザクション処理数が多いほど、数値化処理した後に得られた数値が大きくなる。第１ノード機器に対応するトランザクション割り当て指標が小さいほど、該第１ノード機器に新たなトランザクションを割り当てるマッチング度が高くなることが説明される。

【００５３】

いくつかの実施例において、トランザクション割り当てモードは機器のビジー度に基づく割り当てである。このような場合に、トランザクション割り当てモードによって指示された決定方式に基づき、第１ノード機器に対応するトランザクション割り当て指標を決定する方式としては、該第１ノード機器の機器リソース使用率に基づいて該第１ノード機器に対応するトランザクション割り当て指標を決定する。

【００５４】

第１ノード機器の機器リソース使用率とは、第１ノード機器が既に使用した機器リソースが総機器リソースを占める割合を指し、例示的には、機器リソースとはＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ、中央プロセッサ）リソースを指す。第１ノード機器の機器リソース使用率が高いほど、該第１ノード機器に新たなトランザクションを割り当てるのに適しないことが説明される。説明する必要があるように、第１ノード機器の機器リソース使用率は該第１ノード機器によってリアルタイムに監視されてトランザクション割り当て機器にフィードバックされてもよく、トランザクション割り当て機器によって自ら監視されて得られてもよいが、本願の実施例はこれを限定しない。

【００５５】

該第１ノード機器の機器リソース使用率に基づいて該第１ノード機器に対応するトランザクション割り当て指標を決定する方式は、該第１ノード機器のトランザクション処理数に基づいて該第１ノード機器に対応するトランザクション割り当て指標を決定する方式を参照すればよく、ここでは詳細な説明を省略する。

【００５６】

いくつかの実施例において、トランザクション割り当てモードはハイブリッドのビジー度に基づく割り当てである。このような場合に、トランザクション割り当てモードによって指示された決定方式に基づき、第１ノード機器に対応するトランザクション割り当て指標を決定する方式としては、第１ノード機器のトランザクション処理数、第１ノード機器

の機器リソース使用率、トランザクション処理数の重み、機器リソース使用率の重み、及び重み調節パラメータに基づいて、第１ノード機器に対応するトランザクション割り当て指標を決定する。

【００５７】

例示的な実施例において、トランザクション処理数の重み、及び機器リソース使用率の重みはトランザクション処理数、及び機器リソース使用率の２つのパラメータのパーセンテージを調節することに用いられ、実測して得ることができ、例示的には、トランザクション処理数の重み、及び機器リソース使用率の重みのデフォルト値はいずれも１である。重み調節パラメータとは、機器リソース使用率とトランザクション処理数の相対割合係数を指し、機器リソース使用率とトランザクション処理数の重み割り当てを調節するために用いられ、実測して得ることができ、例示的には、重み調節パラメータのデフォルト値は０．３３である。

【００５８】

例示的には、p_1を利用して機器リソース使用率の重みを表し、p_2を利用してトランザクション処理数の重みを表し、ｗを利用して重み調節パラメータを表すとすると、第１ノード機器に対応するトランザクション割り当て指標ＱはＱ＝p_1×機器リソース使用率＋p_2×ｗ×トランザクション処理数として表されてもよい。

【００５９】

いくつかの実施例において、ハイブリッドのビジー度に基づく割り当てモードにおいて、トランザクション処理数、及び機器リソース使用率を総合的に考慮することに加えて、さらに他の態様の要素、例えば、処理する必要があるトランザクションのうちの長いトランザクションの数などを考慮することができる。このような場合に、第１ノード機器に対応するトランザクション割り当て指標ＱはＱ＝p_1×機器リソース使用率＋p_2×ｗ×トランザクション処理数＋p_3×他の要素として表されてもよい。ここで、p_3は他の要素の重みを表し、p_3は他の要素のタイプに基づき実測して得ることができ、例示的には、p_3のデフォルト値は１である。第１ノード機器に対応するトランザクション割り当て指標Ｑが小さいほど、該第１ノード機器に新たなトランザクションを割り当てるマッチング度が高くなることが説明される。

【００６０】

説明する必要があるように、以上は第１ノード機器の観点からのみ第１ノード機器に対応するトランザクション割り当て指標を決定する過程を紹介し、上記方式に基づき分散型データベースシステムにおける少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定し、更にステップ２０２を実行することを可能にする。

【００６１】

ステップ２０２において、トランザクション割り当て機器は、少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づいて、少なくとも２つのノード機器のうちターゲットトランザクションの協調ノード機器を決定し、協調ノード機器によってターゲットトランザクションを協調処理する。

【００６２】

ターゲットトランザクションの協調ノード機器とは、少なくとも２つのノード機器のうち新たなトランザクションを割り当てるのに適しているノード機器を指す。ターゲットトランザクションの協調ノード機器はターゲットトランザクションを協調処理することに用いられ、つまり、ターゲットトランザクションの協調ノード機器とはターゲットトランザクションのコーディネーターを指す。例示的には、ターゲットトランザクションを協調処理する過程とは、分散型データベースシステムにおいてターゲットトランザクションを発し、次にターゲットトランザクションのデータノード機器を組織して該ターゲットトランザクションを共同で処理する過程を指す。ターゲットトランザクションのデータノード機器とは、少なくとも２つのノード機器のうちターゲットトランザクションの処理に関与することに用いられるノード機器を指し、つまり、ターゲットトランザクションのデータノード機器とはターゲットトランザクションの関与者を指す。

【００６３】
説明する必要があるように、本願の実施例において言及された協調ノード機器及びデータノード機器はいずれもターゲットトランザクションに対するものであり、異なるトランザクションにとって、協調ノード機器又はデータノード機器は固定されるものではなく、すなわち、同一のノード機器はいくつかのトランザクションにとって協調ノード機器に属するが、別のいくつかのトランザクションにとってはデータノード機器に属する可能性がある。
【００６４】
少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づいて、少なくとも２つのノード機器のうちターゲットトランザクションの協調ノード機器を決定する方式は、トランザクション割り当て指標の表現形式の差異に基づき異なっており、本願の実施例はこれを限定せず、協調ノード機器が現在新たなトランザクションを割り当てるのに適しているノード機器であることを確保することを可能にすればよい。
【００６５】
いくつかの実施例において、トランザクション割り当て指標の表現形式はビジーレベルであり、ビジーレベルはそれぞれ「ビジー」、「部分ビジー」及び「アイドル」である。このような場合に、少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づいて、少なくとも２つのノード機器のうちターゲットトランザクションの協調ノード機器を決定する方式としては、少なくとも２つのノード機器のうち対応するトランザクション割り当て指標が「アイドル」のノード機器を候補ノード機器として用い、候補ノード機器において１つのノード機器をターゲットトランザクションの協調ノード機器として任意に選択する。
【００６６】
例示的には、対応するトランザクション割り当て指標が「アイドル」のノード機器が存在しなければ、少なくとも２つのノード機器のうち対応するトランザクション割り当て指標が「部分ビジー」のノード機器を候補ノード機器として用いて、更に候補ノード機器において１つのノード機器をターゲットトランザクションの協調ノード機器として任意に選択する。例示的には、少なくとも２つのノード機器に対応するトランザクション割り当て指標がいずれも「ビジー」であれば、ターゲットトランザクションの協調ノード機器を決定することを一時的に停止し、参照期間待機した後に少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を再決定し、更にターゲットトランザクションの協調ノード機器を再決定する。参照期間は経験に基づき設定され、例えば、参照期間は１つのトランザクションを完了する実測された平均期間である。
【００６７】
いくつかの実施例において、トランザクション割り当て指標の表現形式は数値であり、且つ１つのノード機器に対応するトランザクション割り当て指標が小さいほど、該１つのノード機器に新たなトランザクションを割り当てるマッチング度が高くなることが説明される。このような場合に、少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づいて、少なくとも２つのノード機器のうちターゲットトランザクションの協調ノード機器を決定する方式としては、少なくとも２つのノード機器のうち前ｓ（ｓは１以上の整数である）個の小さなトランザクション割り当て指標に対応するノード機器を候補ノード機器として用いて、候補ノード機器において１つのノード機器をターゲットトランザクションの協調ノード機器として任意に選択する。ｓの値は経験に基づき設定されるか、又は少なくとも２つのノード機器の総数に基づき柔軟に調整され、本願の実施例はこれを限定せず、例えば、ｓの値は１であり、又はｓの値は３などである。
【００６８】
説明する必要があるように、以上は少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づいて、少なくとも２つのノード機器のうちターゲットトランザクションの協調ノード機器を決定する方式の例示的な記述に過ぎず、本願の実施例はこれに限定されない。例示的には、トランザクション割り当て指標の表現形式が数

値であり、且つ１つのノード機器に対応するトランザクション割り当て指標が大きいほど、該１つのノード機器に新たなトランザクションを割り当てるマッチング度が高くなることが説明される場合については、少なくとも２つのノード機器のうち前ｔ（ｔは１以上の整数である）個の大きなトランザクション割り当て指標に対応するノード機器を候補ノード機器として用い、候補ノード機器において１つのノード機器をターゲットトランザクションの協調ノード機器として任意に選択する。

【００６９】

トランザクション割り当て指標に基づいて決定されたターゲットトランザクションの協調ノード機器は少なくとも２つのノード機器のうち新たなトランザクションを割り当てるのに適しているノード機器であり、更にターゲットトランザクションを該協調ノード機器に割り当てて、該協調ノード機器によって該ターゲットトランザクションを協調処理し、ターゲットトランザクションの処理効率を確保することに有利である。

【００７０】

関連技術において、個々のノード機器は一定数のエリア（ｒｅｇｉｏｎ）にサービスし、個々のノード機器はいずれも該ノード機器によってサービスされるエリアにおけるデータ項目の分布情報を維持し、データ項目の分布情報はデータ項目の記憶位置を指示することに用いられる。この他、トランザクション割り当て機器においてエリアのメタ情報が維持される。このようなアーキテクチャでは、関連技術において、トランザクション割り当て機器は維持されたエリアのメタ情報に基づき、ターゲットトランザクションに関するデータ項目が所在するエリアにサービスすることに用いられるノード機器を決定し、更に該ノード機器によってターゲットトランザクションを独立して処理する。このような方式では、トランザクションの処理効率を大幅に制限し、真の分散型トランザクションをサポートすることができず、良好なトランザクション属性特徴付きのグローバルで一致性のあるマルチ読み取り及び一致性のあるマルチ書き込みの能力を備えない。

【００７１】

本願の実施例においては、ノード機器は特定の固定されたエリアにサービスしなくなり、ノード機器はデータ項目の分布情報を維持しなくなり、トランザクション割り当て機器もエリアのメタ情報を維持しなくなる。例示的には、エリアのメタ情報を分散型データベースシステムにおける共有された記憶システム全体において分布する。このような改良に基づいて、トランザクション割り当て機器はトランザクション割り当て指標に基づいてターゲットトランザクションに適切なノード機器を協調ノード機器として割り当てることを可能にし、トランザクションに関するデータ項目を考慮する必要がなく、データ項目の分布状況を考慮する必要もない。ノード機器はトランザクションにおけるＳＱＬ（Ｓｔｒｕｃｔｕｒｅｄ　Ｑｕｅｒｙ　Ｌａｎｇｕａｇｅ、構造化照会言語）ステートメントのニーズに基づき、共有された記憶システムからデータを自動的に呼び出すことを可能にする。このような方式では、個々のノード機器はいずれも分散化されたノード機器として分散型トランザクションを協調処理することを可能にすることにより、分散型データベースシステムが分散化された分散型トランザクションの処理能力を有する。

【００７２】

１つの可能な実現形態において、ターゲットトランザクションの協調ノード機器を決定するステップの後に、協調ノード機器の機器識別情報を該割り当て要求を発した端末に送信するステップをさらに含み、端末は協調ノード機器の機器識別情報に基づき、ターゲットトランザクションのトランザクション情報を協調ノード機器に送信し、協調ノード機器によってトランザクション情報に基づいてターゲットトランザクションを協調処理することに用いられる。

【００７３】

協調ノード機器の機器識別情報は、該協調ノード機器を一意的に識別することに用いられ、協調ノード機器の機器識別情報を割り当て要求を発した端末に送信することによって、端末にターゲットトランザクションを協調処理することに用いられる協調ノード機器を知らせることを可能にする。端末は機器識別情報に基づきターゲットトランザクションを

協調処理することに用いられる協調ノード機器を知った後、ターゲットトランザクションのトランザクション情報を協調ノード機器に送信する。ターゲットトランザクションのトランザクション情報はターゲットトランザクションの関連処理操作を指示することに用いられ、例示的には、ターゲットトランザクションのトランザクション情報とはＳＱＬステートメントを指す。

【００７４】

１つの可能な実現形態において、端末はターゲットトランザクションのトランザクション情報を協調ノード機器に直接送信し、又は、端末はターゲットトランザクションのトランザクション情報及び協調ノード機器の機器識別情報をゲートウェイサーバに送信し、ゲートウェイサーバによってターゲットトランザクションのトランザクション情報を協調ノード機器に転送する。

【００７５】

協調ノード機器はターゲットトランザクションのトランザクション情報を受信した後、トランザクション情報に基づいてターゲットトランザクションを協調処理する。協調ノード機器はトランザクション情報、たとえば、ＳＱＬステートメントを解析し、トランザクション実行計画を生成し、次に関連するデータノード機器と通信することによってターゲットトランザクションの処理を完了することを可能にする。

【００７６】

例示的な実施例において、ターゲットトランザクションの協調ノード機器はトランザクション割り当て指標に基づき決定され、異なるトランザクションは異なるノード機器を利用して協調処理を行うことを可能にするため、本願の実施例が提供する方法は分散化されたトランザクション処理過程を実現することを可能にする。分散化されたトランザクション処理過程において、複数の分散型トランザクションはそれぞれ複数のノード機器によって協調処理される。ターゲットトランザクションの協調ノード機器がターゲットトランザクションを協調処理する過程において、複数の分散型トランザクションが存在すると、協調ノード機器は他のノード機器と通信を確立して、他のノード機器が他の分散型トランザクションを協調処理する過程において生じたデータ情報を取得し、更に取得したデータ情報に基づきデータ異常又はシリアル化可能な認証を行って、ターゲットトランザクションがトランザクションの一致性を満たすか否かを判断し、トランザクション処理技術が正確であることを確保する。例示的な実施例において、協調ノード機器は他のノード機器から送られたデータ情報を一時データバッファ領域にバッファし、ターゲットトランザクションが終了するとクリーンアップする。

【００７７】

１つの可能な実現形態において、ターゲットトランザクションの割り当て要求を発した端末にとって、該端末は分散型データベースシステムにアクセスした後、他のトランザクションが生じると、個々の他のトランザクションはいずれも該協調ノード機器によって協調処理され、又は、個々の他のトランザクションはいずれもトランザクション割り当て機器がノード機器のトランザクション割り当て指標に基づきリアルタイムに割り当てられた適した協調ノード機器によって協調処理されるが、本願の実施例はこれを限定しない。

【００７８】

ステップ２０３において、協調ノード機器はターゲットトランザクションのトランザクション情報を取得する。

【００７９】

ターゲットトランザクションのトランザクション情報はターゲットトランザクションの作成端末によって協調ノード機器に直接送信されてもよく、ゲートウェイサーバによって協調ノード機器に転送されてもよいが、本願の実施例はこれを限定しない。例示的には、ターゲットトランザクションのトランザクション情報とはターゲットトランザクションを実現することに用いられるＳＱＬステートメントを指す。

【００８０】

１つの可能な実現形態において、協調ノード機器はターゲットトランザクションのトラ

ンザクション情報を取得した後、ターゲットトランザクションを初期化する。ターゲットトランザクションを初期化する段階はトランザクションを確立するスナップショット段階とみなされてもよい。この段階ではグローバルな一致性スナップショットポイントを確立し、グローバルな読み取り一致性を保証することを可能にする。

【００８１】

１つの可能な実現形態において、ターゲットトランザクションを初期化する過程において、協調ノード機器は以下の２つの初期化操作のうちの少なくとも１つを実行することができる。

【００８２】

初期化操作１：協調ノード機器はターゲットトランザクションに１つのグローバルで一意的なトランザクション識別子ＴＩＤを割り当てる。

【００８３】

該トランザクション識別子ＴＩＤは該ターゲットトランザクションを一意的に識別することに用いられる。

【００８４】

初期化操作２：協調ノード機器は第１トランザクションステータスリストにおいてターゲットトランザクションの初期ステータス情報を記録する。

【００８５】

本願の実施例においては、協調ノード機器によって維持されたトランザクションステータスリストを第１トランザクションステータスリストと呼び、該第１トランザクションステータスリストは分散化フレームワーク下でターゲットトランザクションのグローバルステータスを記録することに用いられるグローバルステータスリストである。

【００８６】

例示的な実施例において、第１トランザクションステータスリストにおいて記録されたターゲットトランザクションのステータス情報は、ターゲットトランザクションのトランザクション識別子、ターゲットトランザクションのグローバルトランザクションステータス及びターゲットトランザクションの論理ライフサイクルを含むがこれらに限定されない。ここで、論理ライフサイクルはタイムスタンプ下限、及びタイムスタンプ上限で構成される。論理ライフサイクルのタイムスタンプ下限はターゲットトランザクションの開始タイムスタンプ（Ｂｅｇｉｎｔｉｍｅｓｔａｍｐ、Ｂｔｓ）と呼ばれ、論理ライフサイクルのタイムスタンプ上限はターゲットトランザクションの終了タイムスタンプ（Ｅｎｄｔｉｍｅｓｔａｍｐ、Ｅｔｓ）と呼ばれ、つまり、論理ライフサイクルはタイムスタンプ下限Ｂｔｓ、及びタイムスタンプ上限Ｅｔｓで構成される。

【００８７】

ターゲットトランザクションの初期ステータス情報において、ターゲットトランザクションのトランザクション識別子ＴＩＤは初期化操作１において割り当てられ、ターゲットトランザクションのグローバルトランザクションステータスＳｔａｔｕｓはＧｒｕｎｎｉｎｇ（グローバルに動作している）であり、ターゲットトランザクションの論理ライフサイクルは第１論理ライフサイクルであり、該第１論理ライフサイクルのタイムスタンプ下限Ｂｔｓはグローバルで一意的に逓増するタイムスタンプ値であり、第１論理ライフサイクルのタイムスタンプ上限Ｅｔｓは＋∞である。

【００８８】

例示的な実施例において、第１論理ライフサイクルのタイムスタンプ下限Ｂｔｓ、及びタイムスタンプ上限Ｅｔｓの取得方式としては、シリアル化可能レベルより上の分離レベルに対しては、協調ノード機器はグローバルクロックからタイムスタンプ値を取得し、シリアル化可能レベル、及びより弱い分離レベルに対しては、協調ノード機器はローカルのハイブリッド論理クロック（Ｈｙｂｒｉｄ　Ｌｏｇｉｃａｌ　Ｃｌｏｃｋ、ＨＬＣ）からタイムスタンプ値を取得する。もちろん、いくつかの実施例において、シリアル化可能レベル、及びより弱い分離レベルに対しては、協調ノード機器はグローバルクロックからタイムスタンプ値を取得することによって第１論理ライフサイクルのタイムスタンプ下限Ｂ

ｔｓ、及びタイムスタンプ上限Ｅｔｓを取得することも可能にする。例示的な実施例において、シリアル化可能レベル、及びより弱い分離レベルに対して、ローカルのＨＬＣからタイムスタンプ値を取得する効率が比較的高い。

【００８９】

グローバルクロックとはグローバル論理クロックジェネレータによって生成されたクロックを指し、単調に逓増する特性を備え、形式的にはｗａｌｌｔｉｍｅ（ウォールタイム）であってもよく、自然数Ｎなどであってもよい。例示的には、グローバルクロックは分散型データベースシステムにおけるグローバルタイムスタンプ生成クラスタによって提供される。例示的には、ｓｈａｒｅ－ｄｉｓｋアーキテクチャに基づく分散型データベースシステムにとって、グローバルクロックは分散型データベースシステムにおける記憶システムによってＡＰＩの方式を介して提供される。例示的には、グローバル論理クロックジェネレータはトランザクションの開始タイムスタンプＢｔｓ、トランザクションの終了タイムスタンプＥｔｓに値を付与することを可能にし、さらにＷＡＬ（Ｗｒｉｔｅ　Ａｈｅａｄ　Ｌｏｇｇｉｎｇ、ログ先行書き込み）のグローバルＬＳＮ（Ｌｏｇ　Ｓｅｑｕｅｎｃｅ　Ｎｕｍｂｅｒ、ログシーケンス番号）に値を付与することを可能にする。

【００９０】

いくつかの実施例において、グローバルクロックは１つの論理的概念であり、全システムに統一的な単調逓増値を提供し、物理的形態は１つのグローバルな物理クロックであってもよく、１つのグローバルな論理クロックであってもよい。グローバルクロックの実現形態は、複数種の方式を有してもよい。たとえば、グローバルクロックはＧｏｏｇｌｅ（グーグル）の「Ｔｒｕｅｔｉｍｅ（一種のクロックメカニズム）」メカニズムと類似する１つの分散化された分散型クロックであり、又は、グローバルクロックは複数の冗長ノード（たとえば、一致性プロトコル（Ｐａｘｏｓ／Ｒａｆｔなど）で構造されたクラスタ）を用いるマスタースタンバイシステムが統一的に提供した１つのクロックであり、さらに又は、グローバルクロックは正確な同期メカニズムとノードログアウトメカニズムとを併せ持つ一種のアルゴリズムメカニズムによって提供された１つのクロックである。

【００９１】

例示的な実施例において、１つのトランザクションのＢｔｓ、及びＥｔｓの構成はいずれも８バイトからなる。８バイトは２つの部分に分けられ、第１部分は物理タイムスタンプの値（すなわちＵｎｉｘ（一種のオペレーティングシステム）タイムスタンプ、ミリ秒まで正確にする）であり、グローバル時間（ｇｔｓを用いて表す）を識別することに用いられ、第２部分はあるミリ秒内の単調逓増カウントであり、グローバル時間上の相対時間（すなわち局所時間、ｌｔｓを用いて表す）を識別することに用いられる。例示的には、８バイトにおける上位４４ビットは第１部分であり、このように合計で2^{44}個の符号なし整数を表すことができる。従って理論的には合計で約５５７．８（［数１］）年の物理タイムスタンプを表すことができ、８バイトにおける下位２０ビットは第２部分であり、このように、ミリ秒毎に2^{20}個（約１００万個）のカウントがある。例示的な実施例において、２つの部分のビット数を調整することにより、グローバル時間ｇｔｓ、及び局所時間ｌｔｓで表される範囲を変化させることもできる。

【００９２】

【数１】

$$\frac{2^{44}}{1000\times60\times60\times24\times365}=557.8$$

【００９３】

説明する必要があるように、実際の必要に基づき、１つのトランザクションのＢｔｓ、及びＥｔｓの構成は８バイトよりも多いバイトからなるか、又は８バイトよりも少ないバイトからなってもよい。例示的には、Ｂｔｓ、及びＥｔｓの構成を１０バイトからなるよ

うに調整することにより、局所時間ｌｔｓを増大させ、より大きな並行トランザクションの数に対処する。

【００９４】

例示的には、グローバル時間ｇｔｓ、及び局所時間ｌｔｓの２つの部分で構成された２つのタイムスタンプＴｉ．ｂｔｓ、及びＴｊ．ｂｔｓにとって、Ｔｉ．ｂｔｓ．ｇｔｓ＜Ｔｊ．ｂｔｓ．ｇｔｓ、又はＴｉ．ｂｔｓ．ｇｔｓ＝Ｔｊ．ｂｔｓ．ｇｔｓ且つＴｉ．ｂｔｓ．ｌｔｓ＜Ｔｊ．ｂｔｓ．ｌｔｓとなると、Ｔｉ．ｂｔｓ＜Ｔｊ．ｂｔｓとなるとみなされる。

【００９５】

ステップ２０４において、協調ノード機器はターゲットトランザクションのトランザクション情報に基づいて、データノード機器にデータ読み取り要求を送信し、データノード機器は少なくとも２つのノード機器のうちターゲットトランザクションの処理に関与することに用いられるノード機器である。

【００９６】

協調ノード機器はターゲットトランザクションを初期化した後、ターゲットトランザクションの実行段階を開始し、トランザクションの実行段階はトランザクションセマンティクス実現操作段階とみなされてもよい。

【００９７】

データノード機器は少なくとも２つのノード機器のうちターゲットトランザクションの処理に関与することに用いられるノード機器であり、データノード機器はターゲットトランザクションに関するデータ項目を取得することを可能にし、すなわち本願の実施例におけるデータノード機器はターゲットトランザクションに関連するデータノード機器である。ターゲットトランザクションのトランザクション情報において読み取る必要があるデータの関連情報を携え、協調ノード機器はターゲットトランザクションのトランザクション情報に基づき、データ読み取り要求を生成し、次にデータノード機器にデータ読み取り要求を送信することを可能にする。

【００９８】

１つの可能な実現形態において、データ読み取り要求はＲｅａｄＲｅｑｕｅｓｔＭｅｓｓａｇｅ（読み取り要求メッセージ）として表され、ｒｒｑｍと略称する。

【００９９】

１つの可能な実現形態において、データ読み取り要求はターゲットトランザクションの第１論理ライフサイクル、ターゲットトランザクションのトランザクション識別子及び読み取り計画を携えている。第１論理ライフサイクルはタイムスタンプ下限Ｂｔｓ、及びタイムスタンプ上限Ｅｔｓを利用して表される。読み取り計画とはターゲットトランザクションに対応するデータ読み取り計画を指し、読み取る必要があるデータ項目を指示することに用いられる。例示的な実施例において、トランザクション識別子、タイムスタンプ下限Ｂｔｓ、タイムスタンプ上限Ｅｔｓ、及び読み取り計画はそれぞれｒｒｑｍの４つのフィールドにおいて記録される。

【０１００】

説明する必要があるように、データノード機器の数は１つ又は複数であってもよく、本願の実施例においてデータノード機器の数を具体的に限定しない。データノード機器の数が複数である場合については、異なるデータノード機器に送信されたデータ読み取り要求において携える読み取り計画が異なり、異なるデータノード機器において異なるデータ項目を読み取る必要があるように指示する。

【０１０１】

ステップ２０５において、データノード機器は協調ノード機器から送信されたデータ読み取り要求に基づいて、データ読み取り結果を取得し、データ読み取り結果を協調ノード機器に返信する。

【０１０２】

データノード機器はデータ読み取り要求を取得した後、データ読み取り要求に基づいて

、データ読み取り結果を取得する。１つの可能な実現形態において、データ読み取り要求がターゲットトランザクションの第１論理ライフサイクルを携えている場合については、データノード機器がデータ読み取り要求に基づいて、データ読み取り結果を取得する過程は以下のステップ２０５１～ステップ２０５３を含む。

【０１０３】

ステップ２０５１：第１論理ライフサイクルに基づいて、データ読み取り要求によって指示された読み取り対象のデータ項目の可視バージョンデータを決定する。

【０１０４】

データノード機器はデータ読み取り要求が携えている読み取り計画に基づいて、ターゲットトランザクションが読み取る必要があるデータ項目を決定し、ターゲットトランザクションが読み取る必要があるデータ項目を読み取り対象のデータ項目として用いることを可能にする。読み取り対象のデータ項目の可視バージョンデータとは読み取り対象のデータ項目に対応する各バージョンデータのうちターゲットトランザクションに対して可視なあるバージョンデータを指す。例示的には、データノード機器においてデータバッファ領域が設定され、データバッファ領域において読み取り対象のデータ項目の各バージョンデータが存在すると、データノード機器はデータバッファ領域から読み取り対象のデータ項目の各バージョンデータを直接取得し、データバッファ領域において読み取り対象のデータ項目の各バージョンデータが存在しなければ、データノード機器は共有された記憶システムから読み取り対象のデータ項目の各バージョンデータを取得する。

【０１０５】

例示的な実施例において、データノード機器はデータ読み取り要求を受信した後、先ずローカルトランザクションステータスリスト（Ｌｏｃａｌ　ＴＳ）においてターゲットトランザクションのステータス情報が含まれるか否かをチェックする。ローカルトランザクションステータスリストはデータノード機器によって維持されたトランザクションステータスリストであり、ローカルトランザクションステータスリストにおいて該データノード機器が関与するコミットされていない各トランザクションのステータス情報が記録される。例示的な実施例において、データノード機器はデータ読み取り要求を受信した後、データ読み取り要求が携えているターゲットトランザクションのトランザクション識別子に基づき、ローカルトランザクションステータスリストにおいてターゲットトランザクションのステータス情報が含まれるか否かをチェックし、チェック結果は以下の２つの種類を含む。

【０１０６】

チェック結果１：ローカルトランザクションステータスリストにおいてターゲットトランザクションのステータス情報が含まれていない。

【０１０７】

このような場合に、ローカルトランザクションステータスリストにおいて該ターゲットトランザクションのステータス情報を初期化することができ、すなわちＬｏｃａｌ　ＴＳにおいてターゲットトランザクションに関連する１つのレコードを挿入し、該レコードにおける値はそれぞれ、データ読み取り要求が携えているターゲットトランザクションのトランザクション識別子ｒｒｑｍ．ＴＩＤ、データ読み取り要求が携えているターゲットトランザクションの第１論理ライフサイクルのタイムスタンプ下限ｒｒｑｍ．Ｂｔｓ、データ読み取り要求が携えているターゲットトランザクションの第１論理ライフサイクルのタイムスタンプ上限ｒｒｑｍ．Ｅｔｓ、及びデータ読み取り要求によって指示されたターゲットトランザクションの現在のトランザクションステータスｒｒｑｍ．Ｒｕｎｎｉｎｇである。

【０１０８】

このような場合に、第１論理ライフサイクルに基づいて、データ読み取り要求によって指示された読み取り対象のデータ項目の可視バージョンデータを決定する方式としては、読み取り対象のデータ項目の第１論理ライフサイクルに対する可視バージョンデータを決定する。

【０１０９】
チェック結果２：ローカルトランザクションステータスリストにおいてターゲットトランザクションのステータス情報が含まれる。
【０１１０】
このような場合に、データ読み取り要求を受信する前に、ターゲットトランザクションが既に該データノード機器にアクセスしたことが説明される。このとき、該データノード機器上のターゲットトランザクションのステータス情報を更新することができ、更新方法としては、ターゲットトランザクションの論理ライフサイクルのタイムスタンプ下限Ｔ．Ｂｔｓを、照会されたタイムスタンプ下限Ｔ．Ｂｔｓとデータ読み取り要求において携えているタイムスタンプ下限ｒｒｑｍ．Ｂｔｓ（すなわち第１論理ライフサイクルのタイムスタンプ下限）のうちの最大値に更新し、すなわち、Ｔ．Ｂｔｓ＝ｍａｘ（Ｔ．Ｂｔｓ，ｒｒｑｍ．Ｂｔｓ）とする。この他、さらにターゲットトランザクションの論理ライフサイクルのタイムスタンプ上限Ｔ．Ｅｔｓを、照会されたタイムスタンプ上限Ｔ．Ｅｔｓとデータ読み取り要求において携えているタイムスタンプ上限ｒｒｑｍ．Ｅｔｓ（すなわち第１論理ライフサイクルのタイムスタンプ上限）における最小値に更新し、すなわち、Ｔ．Ｅｔｓ＝ｍｉｎ（Ｔ．Ｅｔｓ，ｒｒｑｍ．Ｅｔｓ）とする。更新されたタイムスタンプ下限、及び更新されたタイムスタンプ上限で構成された論理ライフサイクルを更新された論理ライフサイクルとして用いる。
【０１１１】
このような場合に、第１論理ライフサイクルに基づいて、データ読み取り要求によって指示された読み取り対象のデータ項目の可視バージョンデータを決定する方式としては、第１論理ライフサイクルに基づいて、更新された論理ライフサイクルを決定し、読み取り対象のデータ項目の更新された論理ライフサイクルに対する可視バージョンデータを決定する。
【０１１２】
説明する必要があるように、読み取り対象のデータ項目の第１論理ライフサイクルに対する可視バージョンデータを決定する実現形態は、読み取り対象のデータ項目の更新された論理ライフサイクルに対する可視バージョンデータを決定する実現形態と類似しており、本願の実施例においては、読み取り対象のデータ項目の第１論理ライフサイクルに対する可視バージョンデータを決定することを例として説明する。
【０１１３】
１つの可能な実現形態において、読み取り対象のデータ項目の第１論理ライフサイクルに対する可視バージョンデータを決定する前に、先ず第１論理ライフサイクルに対して正当性検査を行って、第１論理ライフサイクルが有効であるか否かを判断する。例示的には、第１論理ライフサイクルに対して正当性検査を行う方式としては、第１論理ライフサイクルのタイムスタンプ下限が第１論理ライフサイクルのタイムスタンプ上限よりも小さいか否かを検査する。第１論理ライフサイクルのタイムスタンプ下限が第１論理ライフサイクルのタイムスタンプ上限以上であれば、第１論理ライフサイクルが無効であることが説明され、このとき、ローカルトランザクションステータスリストにおけるターゲットトランザクションのトランザクションステータスをＲｕｎｎｉｎｇからＡｂｏｒｔｅｄ（アボート）に更新する。この他、データノード機器は協調ノード機器にＡｂｏｒｔ（アボート）メッセージを携えているデータ読み取り結果を返信する。データ読み取り結果はＲｅａｄＲｅｐｌｙＭｅｓｓａｇｅ（読み取りフィードバックメッセージ）として表され、ｒｒｐｍと略称する。データ読み取り結果がＡｂｏｒｔメッセージが携えている場合については、ｒｒｐｍメッセージにおけるＩｓＡｂｏｒｔフィールドは１に等しく、すなわちｒｒｐｍ．ＩｓＡｂｏｒｔ＝１となる。
【０１１４】
第１論理ライフサイクルのタイムスタンプ下限が第１論理ライフサイクルのタイムスタンプ上限よりも小さい場合は、第１論理ライフサイクルが有効であることが説明され、このとき、読み取り対象のデータ項目の第１論理ライフサイクルに対する可視バージョンデ

ータを決定する操作を実行する。

【０１１５】

１つの可能な実現形態において、読み取り対象のデータ項目の第１論理ライフサイクルに対する可視バージョンデータを決定する過程としては、読み取り対象のデータ項目の最新のバージョンデータの作成タイムスタンプが第１論理ライフサイクルのタイムスタンプ上限よりも小さいことに応答して、該最新のバージョンデータを可視バージョンデータとして用いて、読み取り対象のデータ項目の最新のバージョンデータの作成タイムスタンプが第１論理ライフサイクルのタイムスタンプ上限以上であることに応答して、最初の作成タイムスタンプが第１論理ライフサイクルのタイムスタンプ上限よりも小さいあるバージョンデータを決定するまで、読み取り対象のデータ項目の前のバージョンデータと第１論理ライフサイクルのタイムスタンプ上限とを継続的に比較し、該バージョンデータを可視バージョンデータとして用いる。

【０１１６】

つまり、データノード機器は読み取り対象のデータ項目ｘのある論理ライフサイクルに対する可視バージョンデータを決定する過程において、先ず読み取り対象のデータ項目ｘの最新のバージョンデータからチェックを開始し、もし該論理ライフサイクルのタイムスタンプ上限Ｔ．Ｅｔｓが最新のバージョンデータの作成タイムスタンプＷｔｓよりも大きければ、該最新のバージョンデータは該論理ライフサイクルに対する可視バージョンデータである。そうでなければ、該最新のバージョンデータは該論理ライフサイクルに対する可視バージョンデータではなく、Ｔ．Ｅｔｓ＞Ｗｔｓを満たす最初のあるバージョンデータｘ．ｖを見つけるまで、前のバージョンデータを検索する必要があり、該バージョンデータｘ．ｖを該論理ライフサイクルに対する可視バージョンデータとして用いる。

【０１１７】

１つの可能な実現形態において、可視バージョンデータを決定した後、該可視バージョンデータｘ．ｖを該ターゲットトランザクションの読み取りセットにおいて記憶する。選択的に、ここでの読み取りセットはローカル読み取りセットであってもよく、グローバル読み取りセットであってもよい。本願の実施例においては、該読み取りセットがローカル読み取りセットであることを例として説明しており、グローバル読み取りセットの同期に起因する通信オーバーヘッドを回避することを可能にする。

【０１１８】

例示的な実施例において、１つのトランザクションの読み取りセットにおいて該トランザクションが読み取る必要があるデータ項目の可視バージョンデータが記録される。説明する必要があるように、１つの分散型読み取りトランザクションにとっては、該分散型読み取りトランザクションの読み取りセットはローカル読み取りセット及びグローバル読み取りセットに分割されてもよく、ローカル読み取りセットはデータノード機器上に存在し、グローバル読み取りセットは協調ノード機器上に存在する。もちろん、協調ノード機器はグローバル読み取りセットを各データノード機器上に定期的に同期できることにより、データノード機器上にトランザクションのグローバル読み取りセットを維持させることを可能にする。

【０１１９】

ステップ２０５２：可視バージョンデータの作成タイムスタンプ、及び第１論理ライフサイクルに基づいて、ターゲットトランザクションの第２論理ライフサイクルを決定する。

【０１２０】

可視バージョンデータを決定した後、データノード機器は可視バージョンデータの作成タイムスタンプ、及び第１論理ライフサイクルに基づいて、ターゲットトランザクションの第２論理ライフサイクルを決定する。

【０１２１】

いくつかの実施例において、可視バージョンデータとは第１論理ライフサイクルに対する可視バージョンデータを指す場合、該ステップ２０５２の実現形態としては、可視バー

ジョンデータの作成タイムスタンプ、及び第１論理ライフサイクルに直接基づいて、ターゲットトランザクションの第２論理ライフサイクルを決定する。可視バージョンデータとは第１論理ライフサイクルに基づき決定された更新された論理ライフサイクルに対する可視バージョンデータを指す場合、該ステップ２０５２の実現形態としては、可視バージョンデータの作成タイムスタンプ、及び第１論理ライフサイクルに基づき決定された更新された論理ライフサイクルに基づいて、ターゲットトランザクションの第２論理ライフサイクルを決定する。本願の実施例は、可視バージョンデータの作成タイムスタンプ、及び第１論理ライフサイクルに直接基づいて、ターゲットトランザクションの第２論理ライフサイクルを決定することを例として説明する。

【０１２２】

１つの可能な実現形態において、可視バージョンデータの作成タイムスタンプ、及び第１論理ライフサイクルに直接基づいて、ターゲットトランザクションの第２論理ライフサイクルを決定する方式としては、第１論理ライフサイクルのタイムスタンプ下限を調整し、第１論理ライフサイクルのタイムスタンプ下限を可視バージョンデータｘ．ｖの作成タイムスタンプよりも大きくし、すなわち、Ｔ．Ｂｔｓ＞ｘ．ｖ．Ｗｔｓにして、書き読み異常を除去し、調整した後に得られた論理ライフサイクルを第２論理ライフサイクルとして用いる。

【０１２３】

別の可能な実現形態において、可視バージョンデータが読み取り対象のデータ項目の最新のバージョンデータである場合については、可視バージョンデータの作成タイムスタンプ、及び第１論理ライフサイクルに直接基づいて、ターゲットトランザクションの第２論理ライフサイクルを決定する方式としては、第１論理ライフサイクルのタイムスタンプ下限を調整し、第１論理ライフサイクルのタイムスタンプ下限を可視バージョンデータｘ．ｖの作成タイムスタンプよりも大きくし、すなわち、Ｔ．Ｂｔｓ＞ｘ．ｖ．Ｗｔｓにして、書き読み異常を除去し、可視バージョンデータに対応する書き込み対象のトランザクションがヌルではないことに応答して、第１論理ライフサイクルのタイムスタンプ上限を調整し、第１論理ライフサイクルのタイムスタンプ上限を可視バージョンデータに対応する書き込み対象のトランザクションの論理ライフサイクルのタイムスタンプ下限よりも小さくし、すなわちＴ．Ｅｔｓ＜Ｔ０．Ｂｔｓ（Ｔ０は可視バージョンデータに対応する書き込み対象のトランザクションを表す）にして、読み書き競合を除去し、調整した後に得られた論理ライフサイクルを第２論理ライフサイクルとして用いる。

【０１２４】

可視バージョンデータに対応する書き込み対象のトランザクションＷＴは可視バージョンデータと対応するデータ項目を修正しており、且つ認証に合格したトランザクションである。例示的には、書き込み対象のトランザクションのトランザクション識別子を記録することによって書き込み対象のトランザクションを記録する。いくつかの実施例において、可視バージョンデータが最新のバージョンデータである場合については、ターゲットトランザクションのトランザクション識別子を可視バージョンデータのアクティブトランザクション集合において追加し、可視バージョンデータをターゲットトランザクションのローカル読み取りセットにおいて追加する。

【０１２５】

アクティブトランザクション集合（ＲＴｌｉｓｔ）は該最新のバージョンデータにアクセスしたアクティブトランザクションを記録することに用いられ、読み取りトランザクションリストと呼ばれてもよく、該アクティブトランザクション集合は配列の形式であってもよく、リスト、キュー、スタックなどの形式であってもよく、本願の実施例はアクティブトランザクション集合の形式を具体的に限定せず、ＲＴｌｉｓｔにおける個々の要素は上記最新のバージョンデータを読み取ったトランザクションのトランザクション識別子（ＴＩＤ）であってもよい。

【０１２６】

ステップ２０５３：第２論理ライフサイクル、及び可視バージョンデータを携えている

結果をデータ読み取り結果として用いる。

【０１２７】

第２論理ライフサイクル、及び可視バージョンデータを決定した後、データノード機器は第２論理ライフサイクル、及び可視バージョンデータを携えている結果をデータ読み取り結果として用い、次に第２論理ライフサイクル、及び可視バージョンデータを携えているデータ読み取り結果を協調ノード機器に返信し、協調ノード機器に第２論理ライフサイクル、及び可視バージョンデータを取得させる。例示的には、データ読み取り結果はＲｅａｄＲｅｐｌｙＭｅｓｓａｇｅ（読み取りフィードバックメッセージ）として表され、ｒｒｐｍと略称する。例示的には、第２論理ライフサイクル、及び可視バージョンデータを携えているｒｒｐｍにおいてＢｔｓ、Ｅｔｓ、及びＶａｌｕｅフィールドが含まれ、ここで、ＢｔｓフィールドＥｔｓフィールドはそれぞれ第２論理ライフサイクルのタイムスタンプ下限、及び第２論理ライフサイクルのタイムスタンプ上限を記録し、Ｖａｌｕｅフィールドは可視バージョンデータの値を記録する。

【０１２８】

ステップ２０６において、協調ノード機器はデータノード機器から返信されたデータ読み取り結果がトランザクション認証条件を満たすことに応答して、データノード機器にトランザクション認証要求、及びローカル書き込みセットを送信する。

【０１２９】

データノード機器はデータ読み取り結果を協調ノード機器に返信した後、協調ノード機器はデータ読み取り結果がトランザクション認証条件を満たすか否かを判断し、次にデータ読み取り結果がトランザクション認証条件を満たすと判定した場合、データノード機器にトランザクション認証要求、及びローカル書き込みセットを送信し、データノード機器にターゲットトランザクションを認証させる。

【０１３０】

例示的には、協調ノード機器はデータ読み取り結果がトランザクション認証条件を満たすか否かを判断する過程において、先ずデータ読み取り結果がＡｂｏｒｔ（アボート）メッセージを携えているか否かを判断し、すなわちｒｒｐｍにおけるＩｓＡｂｏｒｔフィールドが１に等しいか否かをチェックする。データ読み取り結果がＡｂｏｒｔメッセージを携えており、すなわちｒｒｐｍ．ＩｓＡｂｏｒｔ＝１とすると、データ読み取り結果がトランザクション認証条件を満たさないとみなされ、このとき、グローバルアボート段階に入る。

【０１３１】

データ読み取り結果がＡｂｏｒｔメッセージを携えていなければ、第１トランザクションステータスリストにおけるターゲットトランザクションの論理ライフサイクルを更新し、更新方式としては、第１論理ライフサイクルのタイムスタンプ下限と第２論理ライフサイクルのタイムスタンプ下限における最大値をターゲットトランザクションの第３論理ライフサイクルのタイムスタンプ下限として用い、第１論理ライフサイクルのタイムスタンプ上限と第２論理ライフサイクルのタイムスタンプ上限における最小値をターゲットトランザクションの第３論理ライフサイクルのタイムスタンプ上限として用いる。すなわち、Ｔ．Ｂｔｓ＝ｍａｘ（Ｔ．Ｂｔｓ，ｒｒｐｍ．Ｂｔｓ）、Ｔ．Ｅｔｓ＝ｍｉｎ（Ｔ．Ｅｔｓ，ｒｒｐｍ．Ｅｔｓ）とし、ここで、括弧内のＴ．Ｂｔｓ、及びＴ．Ｅｔｓはそれぞれ更新前の論理ライフサイクル（すなわち第１論理ライフサイクル）のタイムスタンプ下限、及びタイムスタンプ上限であり、ｒｒｐｍ．Ｂｔｓ、及びｒｒｐｍ．Ｅｔｓはそれぞれデータ読み取り結果が携えている第２論理ライフサイクルのタイムスタンプ下限、及びタイムスタンプ上限である。

【０１３２】

第１トランザクションステータスリストにおけるターゲットトランザクションの論理ライフサイクルを更新した後、第１トランザクションステータスリストにおけるＴ．ＢｔｓがＴ．Ｅｔｓよりも小さいか否かをチェックし、すなわち第３論理ライフサイクルのタイムスタンプ下限が第３論理ライフサイクルのタイムスタンプ上限よりも小さいか否かをチ

ェックして、第３論理ライフサイクルが有効であるか否かを判断する。第３論理ライフサイクルのタイムスタンプ下限が第３論理ライフサイクルのタイムスタンプ上限以上である場合、第３論理ライフサイクルが無効であり、このような場合に、データ読み取り結果がトランザクション認証条件を満たさないとみなされ、グローバルアボート段階に入り、第３論理ライフサイクルのタイムスタンプ下限が第３論理ライフサイクルのタイムスタンプ上限よりも小さい場合、第３論理ライフサイクルが有効であり、このような場合に、データ読み取り結果がトランザクション認証条件を満たすとみなされ、データノード機器に第３論理ライフサイクルを携えているトランザクション認証要求を送信する。

【０１３３】

例示的な実施例において、もし協調ノード機器がターゲットトランザクションのアボートを決めれば、第１トランザクションステータスリストにおけるターゲットトランザクションのグローバルトランザクションステータスをＧａｂｏｒｔｉｎｇ（グローバルにアボートしている）に修正し、関連するサブノード（すなわちデータノード機器）に局所アボートを行うように通知する必要がある。

【０１３４】

例示的な実施例において、トランザクション認証要求を送信する前に、協調ノード機器は第１トランザクションステータスリストにおけるターゲットトランザクションのグローバルトランザクションステータスをＧｖａｌｉｄａｔｉｎｇ（グローバルに認証している）に修正する。例示的には、トランザクション認証要求はＶａｌｉｄａｔｅＲｅｑｕｅｓｔＭｅｓｓａｇｅ（認証要求メッセージ）として表され、ｖｒｍと略称する。例示的には、ｖｒｍにおいてＢｔｓ、及びＥｔｓフィールドが含まれる。ここで、Ｂｔｓフィールド及びＥｔｓフィールドはそれぞれターゲットトランザクションの第１トランザクションステータスリストにおける最新の論理ライフサイクルのタイムスタンプ下限、及びタイムスタンプ上限、すなわち第３論理ライフサイクルのタイムスタンプ下限、及びタイムスタンプ上限を記録する。

【０１３５】

いくつかの実施例において、データノード機器の数は複数であり、このような場合に、個々のデータノード機器はいずれも１つのデータ読み取り結果を返信し、データ読み取り結果がトランザクション認証条件を満たすことは、各データノード機器から返信された各データ読み取り結果がいずれもトランザクション認証条件を満たすことを指す。このような場合に、第３論理ライフサイクルは各データ読み取り結果を総合的に考慮することによって決定された論理ライフサイクルである。

【０１３６】

いくつかの実施例において、全部の所要のデータを読み取り且つ更新をローカルの内部メモリにおいて書き込んだ後、トランザクション認証条件を満たすとみなされる。つまり、協調ノード機器は第３論理ライフサイクルが有効であることに応答して、且つターゲットトランザクションのグローバル書き込みセットをローカルの内部メモリにおいて記憶し、データノード機器にトランザクション認証要求を送信する。ターゲットトランザクションのグローバル書き込みセットは端末によって生成されて協調ノード機器に伝送されるか、又は協調ノード機器によって自ら生成され、本願の実施例はこれを限定しない。

【０１３７】

１つのトランザクションの書き込みセットにおいて該トランザクションが更新する必要があるデータ項目が記憶され、読み取りセット構造と類似しており、同様に内部メモリチェーンテーブル構造を使用してトランザクションの書き込みセットを維持することができる。説明する必要があるように、１つの分散型書き込みトランザクションにとって、該分散型書き込みトランザクションの書き込みセットはローカル書き込みセット及びグローバル書き込みセットに分割されてもよく、ローカル書き込みセットはデータノード機器上に存在し、グローバル書き込みセットは協調ノード機器上に存在する。もちろん、協調ノード機器はグローバル書き込みセットを各データノード機器上に定期的に同期できることにより、データノード機器上にもトランザクションのグローバル書き込みセットを維持する

ことを可能にする。

【０１３８】

ターゲットトランザクションのグローバル書き込みセットを協調ノード機器のローカルの内部メモリに書き込んだ後、協調ノード機器はグローバル書き込みセットに基づいてデータノード機器のローカル書き込みセットを決定して、トランザクション認証要求をローカル書き込みセットとともにデータノード機器に送信する。データノード機器のローカル書き込みセットとはターゲットトランザクションのグローバル書き込みセットのうちデータノード機器によって書き込みを担当する必要がある書き込みセットを指す。

【０１３９】

ターゲットトランザクションの読み取り段階では、通信は主に協調ノード機器と関連するデータノード機器との間で発生し、データの読み取りに成功するたびに２回の通信が必要であり、協調ノード機器はデータ読み取り要求を関連するデータノード機器上に送信し、関連するデータノード機器はデータ読み取り結果を協調ノード機器に返信する。従って、データ読み取り段階では、ｎ（ｎが１よりも大きい整数である）がリモート読み取りの回数であると仮定すると、最大２ｎ回の通信を行う必要があり、最大通信量はｎ×（データ読み取り要求メッセージのサイズ＋データ読み取り結果メッセージのサイズ）として表されてもよい。例示的な実施において、ターゲットトランザクションがある関連するデータノード機器の複数のデータ項目のデータを読み取る必要がある場合、複数のデータ項目のデータのデータ読み取り要求をパッケージ化して送信し、これらのデータを一括して読み取り、通信回数を節約し、データ読み取り効率を向上させる。

【０１４０】

ステップ２０７において、データノード機器は協調ノード機器から送信されたトランザクション認証要求、及びローカル書き込みセットに基づいて、ターゲットトランザクションの認証結果を取得し、ターゲットトランザクションの認証結果を協調ノード機器に返信する。

【０１４１】

データノード機器は協調ノード機器から送信されたトランザクション認証要求、及びローカル書き込みセットを受信した後、ターゲットトランザクションの正当性を認証して、ターゲットトランザクションの認証結果を取得する。この段階はトランザクションをコミットする前のトランザクション正当性認証段階である。

【０１４２】

データノード機器の認証過程はローカル認証過程であり、データノード機器がトランザクション認証要求、及びローカル書き込みセットに基づいて、ターゲットトランザクションの認証結果を取得する過程はデータノード機器がローカル認証操作を実行する過程である。１つの可能な実現形態において、トランザクション認証要求は第３論理ライフサイクルを携えており、第３論理ライフサイクルは協調ノード機器によって第１論理ライフサイクル、及び第２論理ライフサイクルに基づいて決定された有効論理ライフサイクルである。第３論理ライフサイクルは協調ノード機器がトランザクション認証要求を送信する前に維持したターゲットトランザクションの最新の論理ライフサイクルである。

【０１４３】

１つの可能な実現形態において、データノード機器はトランザクション認証要求、及びローカル書き込みセットに基づいて、ターゲットトランザクションの認証結果を取得する過程において、データノード機器は先ずローカルトランザクションステータスリストにおけるターゲットトランザクションＴのステータス情報を更新し、更新方式としては、Ｔ．Ｂｔｓ＝ｍａｘ（Ｔ．Ｂｔｓ，ｖｒｍ．Ｂｔｓ）、Ｔ．Ｅｔｓ＝ｍｉｎ（Ｔ．Ｅｔｓ，ｖｒｍ．Ｅｔｓ）とし、ここで、括弧におけるｖｒｍ．Ｂｔｓ、及びｖｒｍ．Ｅｔｓはそれぞれトランザクション認証要求が携えている第３論理ライフサイクルのタイムスタンプ下限、及びタイムスタンプ上限である。本願の実施例において、トランザクション認証要求を受信する前に、データノード機器のローカルトランザクションステータスリストにおいて維持されたターゲットトランザクションの論理ライフサイクルは第２論理ライフサイク

ルであり、区別を容易にするために、トランザクション認証要求を受信した後且つターゲットトランザクションのステータス情報を更新した後に、ローカルトランザクションステータスリストに維持されたターゲットトランザクションの論理ライフサイクルは第４論理ライフサイクルと呼ばれる。

【０１４４】

つまり、データノード機器は第３論理ライフサイクルのタイムスタンプ下限と第２論理ライフサイクルのタイムスタンプ下限における最大値をターゲットトランザクションの第４論理ライフサイクルの下限として用い、第３論理ライフサイクルのタイムスタンプ上限と第２論理ライフサイクルのタイムスタンプ上限における最小値をターゲットトランザクションの第４論理ライフサイクルのタイムスタンプ上限として用いる。これにより、第４論理ライフサイクルを得る。説明する必要があるように、ここで更新されるのはデータノード機器のローカルトランザクションステータスリストにおいて維持されたターゲットトランザクションの論理ライフサイクルであり、このような更新はトランザクション並行アクセス制御に用いられ、すなわちトランザクションの一致性を確保することに用いることを可能にする。

【０１４５】

例示的な実施例において、シリアル化可能分離レベルに対しては、第４論理ライフサイクルを決定した後、第４論理ライフサイクルのタイムスタンプ下限が第４論理ライフサイクルのタイムスタンプ上限よりも小さいか否かをチェックすることによって、第４論理ライフサイクルが有効であるか否かを認証する。

【０１４６】

第４論理ライフサイクルのタイムスタンプ下限が第４論理ライフサイクルのタイムスタンプ上限以上であることに応答して、第４論理ライフサイクルが無効であることが説明され、このとき、ターゲットトランザクションのローカル認証に合格せず、データノード機器は協調ノード機器にＡｂｏｒｔメッセージを携えている認証結果を返信する。該Ａｂｏｒｔメッセージはグローバルアボートを始めることに用いられる。協調ノード機器にターゲットトランザクションの認証結果を返信する過程は協調ノード機器にローカル認証フィードバックメッセージｌｖｍを送信する過程とみなしてもよく、ターゲットトランザクションの認証結果がＡｂｏｒｔメッセージを携えている認証結果である場合については、ローカル認証フィードバックメッセージｌｖｍにおけるＩｓＡｂｏｒｔフィールドは１に等しく、すなわちｌｖｍ．ＩｓＡｂｏｒｔ＝１とする。

【０１４７】

第４論理ライフサイクルのタイムスタンプ下限が第４論理ライフサイクルのタイムスタンプ上限よりも小さいことに応答して、第４論理ライフサイクルが有効であることが説明され、このような場合に、ローカル書き込みセットに対応する各書き込み対象のデータ項目の読み取りトランザクション関連情報及び第４論理ライフサイクルに基づいて、ターゲットトランザクションの第５論理ライフサイクルを決定する。第５論理ライフサイクルとはローカル書き込みセットにおける各書き込み対象のデータ項目に対して読み書き競合認証を行う過程において更新して得られた論理ライフサイクルを指す。

【０１４８】

１つの可能な実現形態において、１つの書き込み対象のデータ項目の読み取りトランザクション関連情報は該１つの書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプ、及び該１つの書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプのうちの少なくとも１つを含む。ここで、１つの書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプ（Ｒｔｓと記される）は該１つの書き込み対象のデータ項目を読み取った各読み取りトランザクションの論理コミットタイムスタンプにおける最大値を指示することに用いられ、１つの書き込み対象のデータ項目のターゲット読み取りトランザクションは該１つの書き込み対象のデータ項目に対応するローカル認証に合格するか、又はコミット段階にある読み取りトランザクションであり、ターゲット読み取りトランザクションの終了タイムスタンプはターゲット読み

取りトランザクションの論理ライフサイクルのタイムスタンプ上限である。
【０１４９】
例示的な実施例において、１つの書き込み対象のデータ項目のターゲット読み取りトランザクションは該１つの書き込み対象のデータ項目に対応するアクティブトランザクション集合におけるローカル認証に合格するか、又はコミット段階にある読み取りトランザクションである。該１つの書き込み対象のデータ項目に対応するアクティブトランザクション集合における各読み取りトランザクションのトランザクションステータスを検出することによって、該１つの書き込み対象のデータ項目のターゲット読み取りトランザクションを決定することができる。
【０１５０】
１つの可能な実現形態において、１つの書き込み対象のデータ項目の読み取りトランザクション関連情報の３つの種類の異なる場合に、ローカル書き込みセットに対応する各書き込み対象のデータ項目の読み取りトランザクション関連情報及び第４論理ライフサイクルに基づいて、ターゲットトランザクションの第５論理ライフサイクルを決定する過程も異なる。
【０１５１】
場合１：１つの書き込み対象のデータ項目の読み取りトランザクション関連情報は該１つの書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプを含む。
【０１５２】
このような場合１に、ローカル書き込みセットに対応する各書き込み対象のデータ項目の読み取りトランザクション関連情報及び第４論理ライフサイクルに基づいて、ターゲットトランザクションの第５論理ライフサイクルを決定する過程としては、各書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプ、及び第４論理ライフサイクルに基づいて、ターゲットトランザクションの第５論理ライフサイクルを決定し、ここで、第５論理ライフサイクルのタイムスタンプ下限は各書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプにおける最大値よりも大きい。
【０１５３】
第４論理ライフサイクルは第５論理ライフサイクルを決定する前に、データノード機器のローカルトランザクションステータスリストにおいて維持されたターゲットトランザクションの最新の論理ライフサイクルである。１つの可能な実現形態において、各書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプ、及び第４論理ライフサイクルに基づいて、ターゲットトランザクションの第５論理ライフサイクルを決定する方式としては、各書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプに基づいて第４論理ライフサイクルのタイムスタンプ下限を調整し、調整した後に得られた論理ライフサイクルを第５論理ライフサイクルとして用いる。
【０１５４】
例示的には、各書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプに基づいて第４論理ライフサイクルのタイムスタンプ下限を調整する方式としては、調整されたタイムスタンプ下限Ｔ．Ｂｔｓ＝ｍａｘ（Ｔ．Ｂｔｓ，ｙ．Ｒｔｓ＋１）とし、ここで、括弧内のＴ．Ｂｔｓは第４論理ライフサイクルのタイムスタンプ下限を表し、ｙ．Ｒｔｓは各書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプにおける最大値を表し、数値１は得られた第５論理ライフサイクルのタイムスタンプ下限が各書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプにおける最大値よりも大きいことを確保することに用いられる。
【０１５５】
いくつかの実施例において、データノード機器はローカル書き込みセットを受信した後に、先ずローカル書き込みセットに対応する各書き込み対象のデータ項目の書き込み対象のトランザクションＷＴがヌルであるか否かを検出し、ある該書き込み対象のデータ項目の書き込み対象のトランザクションＷＴがヌルではなければ、他のトランザクションが該書き込み対象のデータ項目を修正しており、且つ該トランザクションが既に認証段階に入

ったことが説明される。このとき、ターゲットトランザクションをアボートして書き書き競合を除去する必要があり、すなわち協調ノード機器にＡｂｏｒｔメッセージを携えている認証結果を返信する。各書き込み対象のデータ項目の書き込み対象のトランザクションＷＴがいずれもヌルであれば、ターゲットトランザクションのトランザクション識別子の値を各書き込み対象のデータ項目の書き込み対象のトランザクションＷＴに付与し、認証段階に入ったターゲットトランザクションが各書き込み対象のデータ項目を修正する必要があることが表される。実現では、ロックされていないＣＡＳ（Ｃｏｍｐａｒｅ ａｎｄ Ｓｗａｐ、比較と交換）技術を使用して書き込み対象のデータ項目ｙの書き込み対象のトランザクションＷＴに値を付与し、性能を向上させ、又は、先ず書き込み対象のデータ項目ｙの書き込み対象のトランザクションＷＴをロックし、他の並行トランザクションがｙを並行修正することを防止し、次にロックされた書き込み対象のトランザクションＷＴに値を付与する。例示的には、書き込み対象のデータ項目ｙ上にアドバイザリーロックを印加し、該アドバイザリーロックは書き込み対象のデータ項目ｙの書き込み対象のトランザクションＷＴに対する修正操作の相互排他を指示することに用いられる。

【０１５６】

場合２：１つの書き込み対象のデータ項目の読み取りトランザクション関連情報は該１つの書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプを含む。

【０１５７】

このような場合２に、ローカル書き込みセットに対応する各書き込み対象のデータ項目の読み取りトランザクション関連情報及び第４論理ライフサイクルに基づいて、ターゲットトランザクションの第５論理ライフサイクルを決定する過程としては、各書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプ、及び第４論理ライフサイクルに基づいて、ターゲットトランザクションの第５論理ライフサイクルを決定し、ここで、第５論理ライフサイクルのタイムスタンプ下限は各書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプにおける最大値よりも大きい。

【０１５８】

１つの可能な実現形態において、各書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプ、及び第４論理ライフサイクルに基づいて、ターゲットトランザクションの第５論理ライフサイクルを決定する方式としては、各書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプに基づいて第４論理ライフサイクルのタイムスタンプ下限を調整し、調整した後に得られた論理ライフサイクルを第５論理ライフサイクルとして用いる。例示的には、各書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプに基づいて第４論理ライフサイクルのタイムスタンプ下限を調整する方式としては、調整されたタイムスタンプ下限Ｔ．Ｂｔｓ＝ｍａｘ（Ｔ．Ｂｔｓ，Ｔ１．Ｅｔｓ＋１）とし、ここで、括弧内のＴ．Ｂｔｓは第４論理ライフサイクルのタイムスタンプ下限を表し、Ｔ１．Ｅｔｓは各書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプにおける最大値を表し、数値１は得られた第５論理ライフサイクルのタイムスタンプ下限が各書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプにおける最大値よりも大きいことを確保することに用いられる。

【０１５９】

説明する必要があるように、１つの書き込み対象のデータ項目のターゲット読み取りトランザクションの数は１つ又は複数である可能性があり、１つの書き込み対象のデータ項目のターゲット読み取りトランザクションの数が複数である場合については、上記Ｔ１．Ｅｔｓとは全部の書き込み対象のデータ項目の全部のターゲット読み取りトランザクションの終了タイムスタンプにおける最大値を指す。

【０１６０】

このような方式に基づいてターゲットトランザクションの書き込み操作の発生をターゲ

ット読み取りトランザクションの読み取り操作の後に遅延させて、読み書き競合を回避することを可能にする。

【０１６１】

場合３：１つの書き込み対象のデータ項目の読み取りトランザクション関連情報は該１つの書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプ、及び該１つの書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプを含む。

【０１６２】

このような場合３に、ローカル書き込みセットに対応する各書き込み対象のデータ項目の読み取りトランザクション関連情報及び第４論理ライフサイクルに基づいて、ターゲットトランザクションの第５論理ライフサイクルを決定する過程としては、各書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプ、及び各書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプに基づいて第４論理ライフサイクルを連続的に２回調整し、２回調整した後に得られた論理ライフサイクルをターゲットトランザクションの第５論理ライフサイクルとして用いる。本願の実施例は２回調整の前後順序を限定せず、例示的には、先ず各書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプに基づいて第４論理ライフサイクルを調整し、次に各書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプに基づいて１回調整した後に得られた論理ライフサイクルを調整する。もちろん、いくつかの実施例において、さらに先ず各書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプに基づいて第４論理ライフサイクルを調整し、次に各書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプに基づいて１回調整した後に得られた論理ライフサイクルを調整することができる。

【０１６３】

例示的には、２回調整した後に得られた論理ライフサイクルを第５論理ライフサイクルとして用いる場合については、１回調整した後に論理ライフサイクルを得た後、シリアル化可能分離レベルであると、先ず得られた論理ライフサイクルのタイムスタンプ下限がタイムスタンプ上限よりも小さいか否かを認証し、イエスであれば、次回の調整を継続し、ノーであれば、直接ローカル認証が失敗するとみなされ、協調ノード機器にＡｂｏｒｔメッセージを携えている認証結果を返信する。

【０１６４】

第５論理ライフサイクルを得た後、第５論理ライフサイクルのタイムスタンプ下限が第５論理ライフサイクルのタイムスタンプ上限よりも小さいか否かを認証することによって、第５論理ライフサイクルが有効であるか否かを判断する。第５論理ライフサイクルが有効であることに応答して、認証に合格したことを指示することに用いられる認証結果をターゲットトランザクションの認証結果として用い、第５論理ライフサイクルが無効であることに応答して、認証に合格しなかったことを指示することに用いられる認証結果をターゲットトランザクションの認証結果として用いる。認証に合格したことを指示することに用いられる認証結果をターゲットトランザクションの認証結果として用いる場合については、データノード機器のローカル認証フィードバックメッセージｌｖｍにおいてターゲットトランザクションのデータノード機器上で得られた最新の論理ライフサイクル（すなわち第５論理ライフサイクル）のタイムスタンプ下限Ｂｔｓ、及びタイムスタンプ上限Ｅｔｓが記録される。例示的には、認証に合格しなかったことを指示することに用いられる認証結果はＡｂｏｒｔメッセージを携えている認証結果である。

【０１６５】

例示的な実施例において、第５論理ライフサイクルが有効であると判定した場合、ターゲットトランザクションのローカル認証に合格するとみなされ、データノード機器はローカルトランザクションステータスリストにおけるターゲットトランザクションのステータス情報を更新し、ターゲットトランザクションのトランザクションステータスをＶａｌｉｄａｔｅｄ（認証に合格する）に更新し、すなわち、Ｔ．Ｓｔａｔｕｓ＝Ｖａｌｉｄａｔ

ｅｄとする。例示的な実施例において、ターゲットトランザクションのローカル認証に合格すると判定した後、データノード機器は書き込み対象のデータ項目の更新値に基づき、書き込み対象のデータ項目の新たなバージョンデータを作成する。例示的な実施例において、作成された新たなバージョンデータに該新たなバージョンデータがグローバルにコミットされていないことを指示することに用いられる第１マークを設定する。第１マークを有する新たなバージョンデータは外部に対して不可視である。

【０１６６】

説明する必要があるように、ターゲットトランザクションのデータノード機器におけるローカル認証に合格しなければ、データノード機器のローカルトランザクションステータスリストにおけるターゲットトランザクションのトランザクションステータスをＡｂｏｒｔｅｄ（アボート）に更新し、すなわち、Ｔ．Ｓｔａｔｕｓ＝Ａｂｏｒｔｅｄとする必要がある。

【０１６７】

１つの可能な実現形態において、１つの書き込み対象のデータ項目のアクティブトランザクション集合において、ターゲット読み取りトランザクションを含むことに加えて、さらに動作している読み取りトランザクションを含み、ターゲットトランザクションの第５論理ライフサイクルに基づき動作している読み取りトランザクションの論理ライフサイクルを調整する必要があり、動作している読み取りトランザクションにターゲットトランザクションが新たに書き込んだデータを読ませず、それにより読み書き競合現象を回避し、トランザクションが正確に実行されることを確保する。例示的には、動作している読み取りトランザクションとはアクティブトランザクション集合におけるトランザクションステータスがＲｕｎｎｉｎｇ（動作している）のトランザクションを指す。動作している読み取りトランザクションの論理ライフサイクルを調整する方式としては、動作している読み取りトランザクションの論理ライフサイクルのタイムスタンプ上限をターゲットトランザクションの第５論理ライフサイクルのタイムスタンプ下限よりも小さくする。動作している読み取りトランザクションがＴ２であると仮定すると、更新方式としては、Ｔ２．Ｅｔｓ＝ｍｉｎ（Ｔ２．Ｅｔｓ，Ｔ．Ｂｔｓ－１）とする。ある動作している読み取りトランザクションの更新された論理ライフサイクルのタイムスタンプ下限がタイムスタンプ上限以上であると、該動作している読み取りトランザクションにアボートすべきであるように通知する。

【０１６８】

上記トランザクション認証段階から分かるように、ターゲットトランザクションの認証過程において、通信は主に協調ノード機器と関連するデータノード機器との間で発生し、通信は主に、協調ノード機器が個々の関連するデータノード機器にトランザクション認証要求、及びローカル書き込みセットを送信するステップと、関連するデータノード機器が認証結果を協調ノード機器にフィードバックするステップと、を含む。従って、ターゲットトランザクションの認証段階では、ｍ（ｍは１以上の整数である）がターゲットトランザクションＴに関連するデータノード機器の数であると仮定すると、最大２ｍ回の通信を行う必要があり、最大通信量はｍ×（トランザクション認証要求メッセージのサイズ＋認証結果メッセージのサイズ）＋グローバル書き込みセットのサイズとして表されてもよい。

【０１６９】

ステップ２０８において、協調ノード機器はデータノード機器から返信されたターゲットトランザクションの認証結果に基づいて、ターゲットトランザクションの処理命令を決定し、データノード機器に処理命令を送信し、処理命令はコミット命令、又はアボート命令である。

【０１７０】

協調ノード機器はデータノード機器から返信された認証結果を受信した後、受信した認証結果に基づきターゲットトランザクションがグローバル認証に合格するか否かを判断し、更にターゲットトランザクションの処理命令を決定し、データノード機器に処理命令を

送信する。ここで、処理命令はコミット命令、又はアボート命令である。

【０１７１】

１つの可能な実現形態において、データノード機器の数は１つ又は複数であり、データノード機器の数が複数である場合、個々のデータノード機器はいずれも１つの認証結果を返信する。

【０１７２】

データノード機器の数が少なくとも２つである場合について、データノード機器から返信されたターゲットトランザクションの認証結果に基づいて、ターゲットトランザクションの処理命令を決定する過程としては、少なくとも２つのデータノード機器から返信された少なくとも２つの認証結果において認証に合格しなかったことを指示することに用いられる認証結果が存在することに応答して、アボート命令をターゲットトランザクションの処理命令として用いる。少なくとも２つのデータノード機器から返信された少なくとも２つの認証結果はいずれも認証に合格したことを指示することに応答して、少なくとも２つの認証結果が携えている論理ライフサイクルの共通部分をターゲット論理ライフサイクルとして用いて、ターゲット論理ライフサイクルが有効であることに応答して、コミット命令をターゲットトランザクションの処理命令として用いて、ターゲット論理ライフサイクルが無効であることに応答して、アボート命令をターゲットトランザクションの処理命令として用いる。

【０１７３】

例示的な実施例において、認証に合格しなかったことを指示することに用いられる認証結果はＡｂｏｒｔメッセージを携えている認証結果であり、ある認証結果がＡｂｏｒｔメッセージを携えているのではなく、論理ライフサイクル（すなわちステップ２０７において決定された第５論理ライフサイクル）を携えていれば、該認証結果は認証に合格したことを指示する。つまり、協調ノード機器は受信された認証結果に基づきターゲットトランザクションがグローバル認証に合格することを可能にするか否かを判断する過程において、受信された認証結果においてＡｂｏｒｔメッセージを携えている少なくとも１つの認証結果が存在し、すなわちＩｓＡｂｏｒｔフィールドが１に等しいｌｖｍが存在すると、ターゲットトランザクションが全部のローカル認証に合格しないことを表明し、ターゲットトランザクションのグローバル認証に合格せず、ターゲットトランザクションがグローバルアボートを行う必要がある。このような場合に、アボート命令をターゲットトランザクションの処理命令として用いる。協調ノード機器は第１トランザクションステータスリストにおけるターゲットトランザクションのグローバルトランザクションステータスをＧａｂｏｒｔｉｎｇ（グローバルにアボートしている）に更新する。協調ノード機器はデータノード機器にアボート命令を送信して、データノード機器にローカルアボートを行うように通知する。例示的には、処理命令はコミット／アボートメッセージｃｏａｒｍの書き込みによって送信され、処理命令がアボート命令である場合、ｃｏａｒｍにおけるＩｓＡｂｏｒｔフィールドは１に等しく、すなわちｃｏａｒｍ．ＩｓＡｂｏｒｔ＝１とする。

【０１７４】

受信された認証結果においてＡｂｏｒｔメッセージを携えている認証結果が存在しないか、又は受信された認証結果がいずれも論理ライフサイクルを携えていれば、ターゲットトランザクションが全部のローカル認証に合格することが説明される。このような場合に、協調ノード機器は受信された各認証結果が携えている論理ライフサイクルの共通部分を計算して、ターゲット論理ライフサイクルを得る。ターゲット論理ライフサイクルのタイムスタンプ下限がターゲット論理ライフサイクルのタイムスタンプ上限以上であれば、ターゲット論理ライフサイクルが無効であることが説明され、ターゲットトランザクションのグローバル認証に合格しなかったと判定し、ターゲットトランザクションがグローバルアボートを行う必要があり、協調ノード機器はアボート命令をターゲットトランザクションの処理命令として用いる。この他、協調ノード機器はさらに第１トランザクションステータスリストにおけるターゲットトランザクションのグローバルトランザクションステータスをＧａｂｏｒｔｉｎｇ（グローバルにアボートしている）に更新し、協調ノード機器

はデータノード機器にアボート命令を送信して、データノード機器にローカルアボートを行うように通知する。

【０１７５】

ターゲット論理ライフサイクルのタイムスタンプ下限が、ターゲット論理ライフサイクルのタイムスタンプ上限よりも小さければ、ターゲット論理ライフサイクルが有効であることが説明され、ターゲットトランザクションのグローバル認証に合格したと判定し、協調ノード機器はターゲット論理ライフサイクルから１つのタイムスタンプをランダムに選択してターゲットトランザクションの論理コミットタイムスタンプＣｔｓに値を付与し、例えば、ターゲット論理ライフサイクルのタイムスタンプ下限をターゲットトランザクションの論理コミットタイムスタンプとして選択する。

【０１７６】

論理コミットタイムスタンプを決定した後、協調ノード機器は第１トランザクションステータスリストにおけるターゲットトランザクションＴの論理ライフサイクルのタイムスタンプ下限、及び論理ライフサイクルのタイムスタンプ上限をいずれも論理コミットタイムスタンプに更新し、すなわち、Ｔ．Ｂｔｓ＝Ｔ．Ｅｔｓ＝Ｔ．Ｃｔｓとする。この他、第１トランザクションステータスリストにおけるターゲットトランザクションのグローバルトランザクションステータスをＧｃｏｍｍｉｔｔｅｄ（グローバル認証に合格する）に更新し、グローバルタイムスタンプ生成クラスタがターゲットトランザクションにグローバルコミットタイムスタンプを割り当てるように要求し、第１トランザクションステータスリストにおけるターゲットトランザクションのグローバルコミットタイムスタンプＧｔｓフィールドにおいて記録する。この他、協調ノード機器はコミット命令をターゲットトランザクションの処理命令として用い、データノード機器にコミット命令を送信して、データノード機器にターゲットトランザクションをコミットするように通知する。例示的には、処理命令がコミット／アボートメッセージｃｏａｒｍの書き込みによって送信される場合については、処理命令がコミット命令である場合、ｃｏａｒｍにおけるＩｓＡｂｏｒｔフィールドは０に等しく、すなわちｃｏａｒｍ．ＩｓＡｂｏｒｔ＝０とし、ｃｏａｒｍにおけるＣｔｓ、及びＧｔｓフィールドにおいてそれぞれターゲットトランザクションの論理コミットタイムスタンプ、及びターゲットトランザクションのグローバルコミットタイムスタンプが記録される。

【０１７７】

ステップ２０９において、データノード機器は協調ノード機器から送信されたターゲットトランザクションの処理命令を受信したことに応答して、処理命令を実行し、処理命令はコミット命令、又はアボート命令である。

【０１７８】

データノード機器は処理命令を受信した後、処理命令を実行する。データノード機器が処理命令を実行する段階はトランザクションコミット、又はアボート操作終了段階である。

【０１７９】

処理命令がコミット命令である場合、ターゲットトランザクションのグローバル認証に合格し、コミット段階に入ることが説明され、すなわちターゲットトランザクションのデータに対する更新をデータベースにおいて永続化し、かついくつかの後続のクリーンアップ作業を行う。例示的な実施例において、データノード機器は協調ノード機器から送信されたコミット命令を受信した後に、以下の操作Ａ～操作Ｅを実行することができる。

【０１８０】

操作Ａ：ターゲットトランザクションのローカル読み取りセットに対応する個々のデータ項目ｘに対しては、データ項目ｘの最大読み取りトランザクションタイムスタンプＲｔｓを修正し、データ項目ｘの最大読み取りトランザクションタイムスタンプＲｔｓをターゲットトランザクションの論理コミットタイムスタンプＣｔｓよりも大きく又は等しくし、すなわち、ｘ．Ｒｔｓ＝ｍａｘ（ｘ．Ｒｔｓ，Ｔ．Ｃｔｓ）とし、ターゲットトランザクションのトランザクション識別子ＴＩＤを該データ項目ｘのアクティブトランザクショ

ンリストＲＴｌｉｓｔから削除する。
【０１８１】
操作Ｂ：ターゲットトランザクションのローカル書き込みセットに対応する個々のデータ項目ｙに対しては、以下の操作を実行する。ａ）データ項目ｙのオリジナルの作成タイムスタンプＷｔｓをターゲットトランザクションの論理コミットタイムスタンプＴ．Ｃｔｓに修正する。ｂ）データ項目ｙの最大読み取りトランザクションタイムスタンプをオリジナルの最大読み取りトランザクションタイムスタンプとターゲットトランザクションの論理コミットタイムスタンプにおける最大値に更新し、すなわち、ｙ．Ｒｔｓ＝ｍａｘ（ｙ．Ｒｔｓ，Ｔ．Ｃｔｓ）とする。ｃ）データ項目ｙをデータベースにおいて永続化し、且つデータ項目ｙのマークを第１マークから第２マークに修正し、第２マークは外部に対して可視であることを指示することに用いられる。ｄ）データ項目ｙのアクティブトランザクションリストＲＴｌｉｓｔのコンテンツをクリアする。ｅ）データ項目ｙの書き込み対象のトランザクションＷＴのコンテンツをクリアする。
【０１８２】
操作Ｃ：ターゲットトランザクションのローカル読み取りセット及びローカル書き込みセットをクリアする。
【０１８３】
操作Ｄ：ローカルトランザクションステータスリストにおけるターゲットトランザクションの論理ライフサイクルのタイムスタンプ下限、及びタイムスタンプ上限をいずれもターゲットトランザクションの論理コミットタイムスタンプに更新し、すなわち、Ｔ．Ｂｔｓ＝Ｔ．Ｅｔｓ＝Ｔ．Ｃｔｓとする。ローカルトランザクションステータスリストにおけるターゲットトランザクションのトランザクションステータスをＣｏｍｍｉｔｔｅｄ（コミットが完了する）に更新する。説明する必要があるように、このとき、ローカルトランザクションステータスリストは、トランザクションの一致性を確保することに用いられ、グローバルトランザクションステータスの同期に関する必要がない。
【０１８４】
操作Ｅ：協調ノード機器にコミットに成功したＡＣＫ（Ａｃｋｎｏｗｌｅｄｇｅ　Ｃｈａｒａｃｔｅｒ、確認文字）を返信する。
【０１８５】
協調ノード機器は全部のデータノード機器から返信されたコミットに成功したＡＣＫを受信した後、第１トランザクションステータスリストにおけるターゲットトランザクションのグローバルトランザクションステータスをＧｃｏｍｍｉｔｔｅｄ（グローバルコミットが完了する）に修正し、次に協調ノード機器はデータノード機器にステータス情報クリーンアップ命令を送信し、データノード機器にローカルトランザクションステータスリストからターゲットトランザクションのステータス情報を削除させる。
【０１８６】
処理命令がアボート命令である場合、ターゲットトランザクションのグローバル認証に合格しないことが説明され、グローバルアボート段階に入る必要があり、すなわちターゲットトランザクションをアボートし、かつ相応なクリーンアップ作業を行う。例示的には、クリーンアップ作業のコンテンツは、ターゲットトランザクションのトランザクション識別子ＴＩＤをターゲットトランザクションのローカル読み取りセットに対応する個々のデータ項目ｘのアクティブトランザクションリストＲＴｌｉｓｔから削除することと、ターゲットトランザクションのローカル書き込みセットに対応する個々のデータ項目ｙに対応する新たな作成データをクリーンアップし、且つデータ項目ｙの書き込み対象のトランザクションＷＴのコンテンツをクリアすることと、ターゲットトランザクションのローカル読み取りセット及びローカル書き込みセットをクリアすることと、ローカルトランザクションステータスリストにおけるターゲットトランザクションのトランザクションステータスをＡｂｏｒｔｅｄ（アボートが完了する）に更新することと、協調ノード機器にアボートが完了したＡＣＫを返信することと、を含む。
【０１８７】

協調ノード機器は全部のデータノード機器から返信されたアボートが完了したＡＣＫを受信した後、第１トランザクションステータスリストにおけるターゲットトランザクションのグローバルトランザクションステータスをＧａｂｏｒｔｅｄ（グローバルアボートが完了する）に修正し、次に協調ノード機器はデータノード機器にステータス情報クリーンアップ命令を送信し、データノード機器にローカルトランザクションステータスリストからターゲットトランザクションのステータス情報を削除させる。１つの可能な実現形態において、協調ノード機器はデータノード機器にステータス情報クリーンアップ命令を一括して送信して、通信回数を減少させる。

【０１８８】

上記内容から明らかなように、ターゲットトランザクションのコミット／アボート段階では、通信は主に協調ノード機器と関連するデータノード機器との間で発生し、通信は主に、協調ノード機器が個々の関連するデータノード機器にコミット／アボート命令を送信するステップと、個々の関連するデータノード機器が協調ノード機器に相応なコミット／アボート完了メッセージ（ＡＣＫ）を送信するステップと、を含む。従って、コミット／アボート段階では、最大２ｍ回の通信を行い、通信量のサイズはｍ×（コミット／アボート命令メッセージのサイズ＋コミット／アボート完了メッセージのサイズ）であり、ここでｍ（ｍは１以上の整数である）はターゲットトランザクションＴに関連するデータノード機器の数である。

【０１８９】

説明する必要があるように、本願の実施例はターゲットトランザクションが読み書き操作に関することを例として紹介したが、本願の実施例はこれに限定されず、ターゲットトランザクションが読み取り操作のみに関するか、又は書き込み操作のみに関する場合については、依然として本願の実施例が提供するトランザクション処理方法に基づきトランザクションに対する処理を実現することを可能にするが、本願の実施例は詳細な説明を省略する。

【０１９０】

上記ステップ２０１～ステップ２０９に基づいてトランザクションを処理する過程はトランザクションの分散化処理を実現し、並行トランザクション間の競合操作に起因するデータ異常の問題を解決することを可能にする。実現原理から見ると、本願の実施例が提供するトランザクション処理方法は主にＯＣＣ（Ｏｐｔｉｍｉｓｔｉｃ　Ｃｏｎｃｕｒｒｅｎｃｙ　Ｃｏｎｔｒｏｌ、楽観的並行性制御）のアルゴリズムフレームワークを応用し、ＤＴＡ（Ｄｙｎａｍｉｃ　Ｔｉｍｅｓｔａｍｐ　Ａｌｌｏｃａｔｉｏｎ、動的割り当てタイムスタンプ）アルゴリズムを組み合わせて、ネットワークで伝送されたトランザクションデータ情報を減少させ、分散型トランザクションの認証効率を向上させ、分散型トランザクションの並行処理能力を高める。この他、さらにＭＶＣＣ（Ｍｕｔｉｌ－Ｖｅｒｓｉｏｎ　Ｃｏｎｃｕｒｒｅｎｃｙ　Ｃｏｎｔｒｏｌ、マルチバージョン並行性制御）を組み合わせてロックされていないデータ読み書きを実現し、それにより局所ノード機器の並行処理能力を高める。ここで、ＤＴＡアルゴリズムはＴＯ（Ｔｉｍｅｓｔａｍｐ　Ｏｒｄｅｒｉｎｇ、タイムスタンプオーダリング）アルゴリズムに属し、トランザクションの論理ライフサイクルのタイムスタンプ下限、及びタイムスタンプ上限を動的に調整することができる。

【０１９１】

本願の実施例が提供する方法はデータ記憶フォーマットに影響されず、本願の実施例における分散型データベースシステムはキー値型のデータ記憶フォーマット（ＫＶデータ記憶フォーマット）（たとえば、ＨＢａｓｅデータベースシステムにおけるデータ記憶フォーマット）をサポートするだけでなく、セグメントページ型のデータ記憶フォーマット（たとえば、ＰｏｓｔｇｒｅＳＱＬ及びＭｓＳＱＬ／ＩｎｎｏＤＢデータベースシステムにおけるデータ記憶フォーマット）をサポートする。

【０１９２】

例示的な実施例において、セグメントページ型のデータ記憶フォーマットに対しては、

ノード機器内にデータバッファ領域を確立し、共有された記憶システムから伝送されたデータをバッファし、それによりデータを次回取得する速度を加速させ、バッファフォーマットは下層のデータ記憶フォーマットと一致するように保持される。共有された記憶システムから伝送されたデータは、ローカルのデータバッファ領域にバッファされ、ローカルデータバッファ領域がいっぱいになるか、又はダーティデータを共有された記憶システムにフラッシュバックする必要があるか、又はバッファが無効になる（たとえば、他のノード機器上で同様のデータが修正される）ときまで、トランザクションが終了してもクリーンアップされない。

【０１９３】

トランザクションがコミットされる前に、個々のノード機器は共有された記憶システムにトランザクションログ（たとえば、ＷＡＬログ）を算出し、トランザクションログは共有された記憶システムにＬＳＮ値を請求し、該値はグローバルで一意的で且つ逓増する１つの値である。異なるデータ記憶フォーマットでは、トランザクション処理過程において生じたトランザクションログは異なるフォーマットを有する。例示的には、データ記憶フォーマットがＫＶデータ記憶フォーマットである場合、トランザクションログのフォーマットは図３に示される。

【０１９４】

データベースシステムによって維持された大規模なテーブルで分割された各エリア（Ｒｅｇｉｏｎ）は１つのログファイルを共有し、単一のエリアはログにおいて時間順序に応じて記憶されるが、複数のエリアは完全に時間順序に応じたものではない可能性がある。個々のログの最小ユニットはログキー（ＨＬｏｇＫｅｙ）及びログ編集（ＷＡＬＥｄｉｔ）の２つの部分からなる。ここで、ＨＬｏｇＫｅｙはシーケンス番号（ｓｅｑｕｅｎｃｅｉｄ）、タイムスタンプ（ｔｉｍｅｓｔａｍｐ）、クラスタ番号（ｃｌｕｓｔｅｒ　ｉｄｓ）、エリア名（ｒｅｇｉｏｎ　ｎａｍｅ）及びテーブル名（ｔａｂｌｅ　ｎａｍｅ）などからなり、ＷＡＬＥｄｉｔは一連のキー値ペア（Ｋｅｙ　Ｖａｌｕｅ）からなり、１行におけるすべての列（すなわちすべてのＫｅｙ　Ｖａｌｕｅ）に対する更新操作はいずれも同一のＷＡＬＥｄｉｔオブジェクトにおいて含まれ、これは主に１行複数列に書き込むときの原子性を実現するためである。ｓｅｑｕｅｎｃｅｉｄは、１つの記憶レベルの自己増加シーケンス番号であり、エリアのデータ復旧及び期限切れログのクリアはいずれもそれに依存し、例示的には、ｓｅｑｕｅｎｃｅｉｄとはトランザクションログのＬＳＮ値を指す。

【０１９５】

例示的には、データ記憶フォーマットがセグメントページ型のデータ記憶フォーマットである場合、トランザクションログのフォーマットは図４に示される。個々のＲｅｇｉｏｎは１つのログファイルを共有し、単一のＲｅｇｉｏｎはログにおいて時間順序に応じて記憶され、且つ複数のＲｅｇｉｏｎは完全に時間順序に応じたものではない可能性がある。個々のログの最小ユニットはＨＬｏｇｋｅｙ及びＷＡＬＥｄｉｔの２つの部分からなるのではなく、１つのログレコード（ＸＬｏｇ　Ｒｅｃｏｒｄ）からなる。

【０１９６】

ＸＬｏｇ　Ｒｅｃｏｒｄは２つの部分で構成され、第１部分はヘッダ情報であり、サイズが固定され（たとえば、２４Ｂ（Ｂｙｔｅｓ、バイト）、対応する構造体はＸＬｏｇＲｅｃｏｒｄであり、第２部分はログレコードデータ（ＸＬｏｇ　Ｒｅｃｏｒｄ　ｄａｔａ）である。

【０１９７】

ＸＬｏｇ　Ｒｅｃｏｒｄは記憶されたデータコンテンツに応じて分割され、主に以下の３つの種類に分けられる。

【０１９８】

第１種類：Ｒｅｃｏｒｄ　ｆｏｒ　ｂａｃｋｕｐ　ｂｌｏｃｋ（バックアップブロックレコード）：ｆｕｌｌ－ｗｒｉｔｅ－ｐａｇｅ（全書き込みページ）を記憶するｂｌｏｃｋ（ブロック）であり、このようなタイプのレコードはページ部分の書き込みの問題を解

決するためである。ｃｈｅｃｋｐｏｉｎｔ（検出ポイント）が完了した後にデータページを１回目に修正し、この変更を記録してトランザクションログファイルに書き込むときにページ全体に書き込む（相応な初期化パラメータを設定する必要があり、デフォルトで開いている）。

【０１９９】

第２種類：Ｒｅｃｏｒｄ ｆｏｒ ｔｕｐｌｅ ｄａｔａ ｂｌｏｃｋ（タプルデータブロックレコード）：ページにおけるタプル変更を記憶することに用いられる。

【０２００】

第３種類：Ｒｅｃｏｒｄ ｆｏｒ Ｃｈｅｃｋｐｏｉｎｔ（チェックポイントレコード）：ｃｈｅｃｋｐｏｉｎｔが発生した場合、トランザクションログファイルにおいてｃｈｅｃｋｐｏｉｎｔ情報（ここでＲｅｄｏ ｐｏｉｎｔ（やり直しポイント）を含む）を記録する。

【０２０１】

ＸＬｏｇ Ｒｅｃｏｒｄ ｄａｔａは実際のデータを記憶する場所であり、以下の４つの部分からなる。

【０２０２】

第１部分：０－Ｎ個のＸＬｏｇＲｅｃｏｒｄＢｌｏｃｋＨｅａｄｅｒ（ログレコードブロックヘッダ）であり、各々のＸＬｏｇＲｅｃｏｒｄＢｌｏｃｋＨｅａｄｅｒは１つのｂｌｏｃｋ ｄａｔａ（ブロックデータ）に対応する。ＢＫＰＢＬＯＣＫ＿ＨＡＳ＿ＩＭＡＧＥマークを設定すると、ＸＬｏｇＲｅｃｏｒｄＢｌｏｃｋＨｅａｄｅｒ構造体の後にＸＬｏｇＲｅｃｏｒｄＢｌｏｃｋＩｍａｇｅＨｅａｄｅｒ構造体が続き、ＢＫＰＢＬＯＣＫ＿ＨＡＳ＿ＨＯＬＥ＆ＢＫＰＩＭＡＧＥ＿ＩＳ＿ＣＯＭＰＲＥＳＳＥＤマークを設定すると、ＸＬｏｇＲｅｃｏｒｄＢｌｏｃｋＨｅａｄｅｒ構造体の後にＸＬｏｇＲｅｃｏｒｄＢｌｏｃｋＣｏｍｐｒｅｓｓＨｅａｄｅｒ構造体が続き、ＢＫＰＢＬＯＣＫ＿ＳＡＭＥ＿ＲＥＬマークを設定しなければ、ＸＬｏｇＲｅｃｏｒｄＢｌｏｃｋＨｅａｄｅｒ構造体の後にＲｅｌＦｉｌｅＮｏｄｅが続く。例示的には、ＸＬｏｇＲｅｃｏｒｄＢｌｏｃｋＨｅａｄｅｒ構造体の後にさらにＢｌｏｃｋＮｕｍｂｅｒ（ブロック番号）が続いてもよい。

【０２０３】

第２部分：ＸＬｏｇＲｅｃｏｒｄＤａｔａＨｅａｄｅｒ［Ｓｈｏｒｔ｜Ｌｏｎｇ］（ログレコードデータヘッダ［ショート｜ロング］）であり、データのサイズ＜２５６ Ｂｙｔｅｓとすると、Ｓｈｏｒｔフォーマットを使用し、そうでなければ、Ｌｏｎｇフォーマットを使用する。

【０２０４】

第３部分：ｂｌｏｃｋ ｄａｔａ（ブロックデータ）：ｆｕｌｌ－ｗｒｉｔｅ－ｐａｇｅ ｄａｔａ（全書き込みページデータ）及びｔｕｐｌｅ ｄａｔａ（タプルデータ）。ｆｕｌｌ－ｗｒｉｔｅ－ｐａｇｅ ｄａｔａに対しては、圧縮をイネーブルにすると、データが圧縮記憶され、圧縮された後に該ｐａｇｅに関連するメタデータをＸＬｏｇＲｅｃｏｒｄＢｌｏｃｋＣｏｍｐｒｅｓｓＨｅａｄｅｒ（ログレコードブロック圧縮ヘッダ）に記憶する。

【０２０５】

第４部分：ｍａｉｎ ｄａｔａ（主要データ）：ｃｈｅｃｋｐｏｉｎｔなどのログデータを記録する。

【０２０６】

例示的には、ＸＬｏｇ Ｒｅｃｏｒｄは以下のように定義される。

ヘッダ情報（固定サイズのＸＬｏｇＲｅｃｏｒｄ構造体）

ＸＬｏｇＲｅｃｏｒｄＢｌｏｃｋＨｅａｄｅｒ 構造体

ＸＬｏｇＲｅｃｏｒｄＢｌｏｃｋＨｅａｄｅｒ 構造体

．．．

ＸＬｏｇＲｅｃｏｒｄＤａｔａＨｅａｄｅｒ［Ｓｈｏｒｔ｜Ｌｏｎｇ］ 構造体

ｂｌｏｃｋ ｄａｔａ

ｂｌｏｃｋ ｄａｔａ
・・・
ｍａｉｎ ｄａｔａ

【０２０７】
１つの可能な実現形態において、データ記憶フォーマットがセグメントページ型のデータ記憶フォーマットである場合については、並行トランザクションが異なるノード機器（ＥＳ）上で処理され、且つ同一のページ上の異なるデータ項目を修正する場合、ページレベルの競合が発生してデータ上書きの問題を引き起こすようになる。たとえば、トランザクションＴａはノード機器ＥＳ－１上でＸ＝２のデータ項目を修正し、トランザクションＴｂはノード機器ＥＳ－２上でＸ＝３のデータ項目を修正し、Ｘ＝２とＸ＝３のデータ項目は同一のページ（ｐａｇｅ）上にあり、このとき、トランザクション処理メカニズムは、並行に動作され、トランザクションを同時実行し、トランザクションレベルでデータ異常が存在しない。しかし、ページレベルでは、ＥＳ－１とＥＳ－２のどちらでフラッシュアウトされたトランザクションログを選ぶべきであるかという選択が存在し、これは同一の物理ページに対する修正が両立できないという問題の出現をもたらす。

【０２０８】
セグメントページ型のデータ記憶フォーマットをサポートするトランザクションログにおいて、１つのセグメントページ型のリスト（ｌｉｓｔ）が増加され、本セグメントログにおけるページのアドレス（たとえば、ファイル番号、テーブルスペース番号、ファイルにおける相対オフセットなど）及び個々のページ上で書き込み操作を実行しているトランザクション識別子がマーキングされる。トランザクションログが最下位層に共有された記憶システムにフラッシュアウトされる場合、認証機器はすべての並行トランザクションの共有された記憶システムのトランザクションにおけるｌｉｓｔにコミットされたページが重なるか否かをチェックする。もし重なっていれば、並行トランザクションが同じページを書き込んだことが表明され（書き込むのが同じデータ項目であれば、トランザクション認証段階では、トランザクション競合が存在すると検出しかつ既にアボートによって競合を解消している）、書き込むのが同じページ上の異なるデータ項目であり、このときにページレベルの競合が存在し、データ上書きイベントが発生し、そのうちの１つのノード機器ＥＳに対応するトランザクションをアボートする必要があり、対応するトランザクションがアボートされたノード機器ＥＳのトランザクションログをフラッシュアウトしなくなり、問題の発生を回避する。

【０２０９】
例示的な実施例において、上記ページレベルの競合認証を実行する主体は分散型データベースシステムにおける認証機器である。該認証機器はいずれかのノード機器と同一の物理マシン上にあってもよく、独立した機器であってもよく、本願の実施例はこれを限定しない。

【０２１０】
本願の実施例が提供するトランザクション処理方法に基づいて、分散型データベースシステムが分散型トランザクションをサポートするだけでなく、グローバルで一致性のあるマルチ読み取りを達成することを可能にし、分散化トランザクション処理技術によって性能にも配慮することを可能にする。良好なトランザクション属性特徴付きのグローバルで一致性のあるマルチ読み取り及び一致性のあるマルチ書き込み能力を備える。本願の実施例が提供するトランザクション処理方法に基づいて、ｓｈａｒｅ－ｄｉｓｋアーキテクチャに基づく分散型データベースシステム、たとえば有名なＮｏＳＱＬ（Ｎｏｎ－ｒｅｌａｔｉｏｎａｌ ＳＱＬ、一般的に非リレーショナルデータベースを指す）下でのＨＢａｓｅデータベースシステムに、分散化された分散型トランザクション処理手段を提供できることにより、ＨＢａｓｅと類似するデータベースシステムはエリア間、ノード間の高効率なトランザクション処理能力を備える。

【０２１１】
本願の実施例においては、各ノード機器のそれぞれに対応するトランザクション割り当

て指標に基づきターゲットトランザクションを協調処理することに用いられる協調ノード機器を決定し、トランザクションの割り当て過程はトランザクションに関するデータ項目を考慮する必要がなく、データ項目の分布状況を考慮する必要もない。このような方式に基づいて、個々のノード機器はいずれも分散化された機器としてトランザクションを協調処理することを可能にすることにより、トランザクションをノード間で処理することを可能にし、トランザクションの処理効率を向上させることに有利であり、トランザクション処理の信頼性が比較的高く、データベースシステムのシステム性能を高めることに有利である。

【０２１２】

本願の実施例はトランザクション処理システムを提供し、該トランザクション処理システムは協調ノード機器と、データノード機器とを含み、協調ノード機器は同一の記憶システムを共有する少なくとも２つのノード機器のうちターゲットトランザクションを協調処理することに用いられるノード機器であり、協調ノード機器は少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づき決定され、データノード機器は少なくとも２つのノード機器のうちターゲットトランザクションの処理に関与することに用いられるノード機器であり、

ここで、協調ノード機器は、ターゲットトランザクションのトランザクション情報を取得することと、ターゲットトランザクションのトランザクション情報に基づいて、データノード機器にデータ読み取り要求を送信することと、に用いられ、

データノード機器は、協調ノード機器から送信されたデータ読み取り要求に基づいて、データ読み取り結果を取得し、データ読み取り結果を協調ノード機器に返信することに用いられ、

協調ノード機器はさらに、データノード機器から返信されたデータ読み取り結果がトランザクション認証条件を満たすことに応答して、データノード機器にトランザクション認証要求、及びローカル書き込みセットを送信することに用いられ、

データノード機器はさらに、協調ノード機器から送信されたトランザクション認証要求、及びローカル書き込みセットに基づいて、ターゲットトランザクションの認証結果を取得し、ターゲットトランザクションの認証結果を協調ノード機器に返信することに用いられ、

協調ノード機器はさらに、データノード機器から返信されたターゲットトランザクションの認証結果に基づいて、ターゲットトランザクションの処理命令を決定し、データノード機器に処理命令を送信することに用いられ、処理命令はコミット命令、又はアボート命令であり、

データノード機器はさらに、協調ノード機器から送信されたターゲットトランザクションの処理命令を受信したことに応答して、処理命令を実行することに用いられる。

【０２１３】

１つの可能な実現形態において、データ読み取り結果は第２論理ライフサイクルを携えており、第２論理ライフサイクルはデータノード機器によってデータ読み取り要求が携えているターゲットトランザクションの第１論理ライフサイクルに基づき決定され、第１論理ライフサイクルはタイムスタンプ下限、及びタイムスタンプ上限で構成され、協調ノード機器はさらに、第１論理ライフサイクルのタイムスタンプ下限と第２論理ライフサイクルのタイムスタンプ下限における最大値をターゲットトランザクションの第３論理ライフサイクルのタイムスタンプ下限として用いることと、第１論理ライフサイクルのタイムスタンプ上限と第２論理ライフサイクルのタイムスタンプ上限における最小値をターゲットトランザクションの第３論理ライフサイクルのタイムスタンプ上限として用いることと、第３論理ライフサイクルが有効であることに応答して、データノード機器に第３論理ライフサイクルを携えているトランザクション認証要求を送信することであって、第３論理ライフサイクルが有効であることは、第３論理ライフサイクルのタイムスタンプ下限が第３論理ライフサイクルのタイムスタンプ上限よりも小さいことを指示することに用いられる、ことと、に用いられる。

【０２１４】
１つの可能な実現形態において、データノード機器の数は少なくとも２つであり、協調ノード機器はさらに、少なくとも２つのデータノード機器から返信された少なくとも２つの認証結果において認証に合格しなかったことを指示することに用いられる認証結果が存在することに応答して、アボート命令をターゲットトランザクションの処理命令として用いることと、少なくとも２つのデータノード機器から返信された少なくとも２つの認証結果はいずれも認証に合格したことを指示することに応答して、少なくとも２つの認証結果が携えている論理ライフサイクルの共通部分をターゲット論理ライフサイクルとして用いることと、ターゲット論理ライフサイクルが有効であることに応答して、コミット命令をターゲットトランザクションの処理命令として用いることと、ターゲット論理ライフサイクルが無効であることに応答して、アボート命令をターゲットトランザクションの処理命令として用いることと、に用いられる。
【０２１５】
１つの可能な実現形態において、データ読み取り要求はターゲットトランザクションの第１論理ライフサイクルを携えており、第１論理ライフサイクルはタイムスタンプ下限、及びタイムスタンプ上限で構成され、データノード機器は、第１論理ライフサイクルに基づいて、データ読み取り要求によって指示された読み取り対象のデータ項目の可視バージョンデータを決定することと、可視バージョンデータの作成タイムスタンプ、及び第１論理ライフサイクルに基づいて、ターゲットトランザクションの第２論理ライフサイクルを決定することと、第２論理ライフサイクル、及び可視バージョンデータを携えている結果をデータ読み取り結果として用いることと、に用いられる。
【０２１６】
１つの可能な実現形態において、トランザクション認証要求はターゲットトランザクションの第３論理ライフサイクルを携えており、第３論理ライフサイクルは協調ノード機器によって第１論理ライフサイクル、及び第２論理ライフサイクルに基づいて決定された有効論理ライフサイクルであり、データノード機器はさらに、第３論理ライフサイクルのタイムスタンプ下限と第２論理ライフサイクルのタイムスタンプ下限における最大値をターゲットトランザクションの第４論理ライフサイクルのタイムスタンプ下限として用いることと、第３論理ライフサイクルのタイムスタンプ上限と第２論理ライフサイクルのタイムスタンプ上限における最小値をターゲットトランザクションの第４論理ライフサイクルのタイムスタンプ上限として用いることと、第４論理ライフサイクルが有効であることに応答して、ローカル書き込みセットに対応する各書き込み対象のデータ項目の読み取りトランザクション関連情報及び第４論理ライフサイクルに基づいて、ターゲットトランザクションの第５論理ライフサイクルを決定することと、第５論理ライフサイクルが有効であることに応答して、認証に合格したことを指示することに用いられる認証結果をターゲットトランザクションの認証結果として用いることと、第５論理ライフサイクルが無効であることに応答して、認証に合格しなかったことを指示することに用いられる認証結果をターゲットトランザクションの認証結果として用いることと、に用いられる。
【０２１７】
１つの可能な実現形態において、１つの書き込み対象のデータ項目の読み取りトランザクション関連情報は１つの書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプを含み、１つの書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプは、１つの書き込み対象のデータ項目を読み取った各読み取りトランザクションの論理コミットタイムスタンプにおける最大値を指示することに用いられ、データノード機器はさらに、各書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプ、及び第４論理ライフサイクルに基づいて、ターゲットトランザクションの第５論理ライフサイクルを決定することに用いられ、第５論理ライフサイクルのタイムスタンプ下限は各書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプにおける最大値よりも大きい。
【０２１８】

１つの可能な実現形態において、１つの書き込み対象のデータ項目の読み取りトランザクション関連情報は１つの書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプを含み、ターゲット読み取りトランザクションはローカル認証に合格するか、又はコミット段階にある読み取りトランザクションであり、ターゲット読み取りトランザクションの終了タイムスタンプはターゲット読み取りトランザクションの論理ライフサイクルのタイムスタンプ上限であり、データノード機器はさらに、各書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプ、及び第４論理ライフサイクルに基づいて、ターゲットトランザクションの第５論理ライフサイクルを決定することに用いられ、第５論理ライフサイクルのタイムスタンプ下限は各書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプにおける最大値よりも大きい。

【０２１９】

上記実施例が提供するシステムは方法実施例と同一の構想に属し、その具体的な実現過程については、詳しくは方法実施例を参照すればよいため、ここでは詳細な説明を省略する。

【０２２０】

図５に参照されるように、本願の実施例はトランザクション処理装置を提供し、該装置は、

ターゲットトランザクションの割り当て要求に応答して、同一の記憶システムを共有する少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定することに用いられる第１決定ユニット５０１であって、１つのノード機器に対応するトランザクション割り当て指標は該１つのノード機器に新たなトランザクションを割り当てるマッチング度を指示することに用いられる、第１決定ユニット５０１と、

少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づいて、少なくとも２つのノード機器のうちターゲットトランザクションの協調ノード機器を決定し、協調ノード機器によってターゲットトランザクションを協調処理することに用いられる第２決定ユニット５０２と、を含む。

【０２２１】

１つの可能な実現形態において、第１決定ユニット５０１は、トランザクション割り当てモードを決定することであって、トランザクション割り当てモードはトランザクションのビジー度に基づく割り当て、機器のビジー度に基づく割り当て、及びハイブリッドのビジー度に基づく割り当てのうちのいずれか１つを含む、ことと、トランザクション割り当てモードによって指示された決定方式に基づき、少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定することと、に用いられる。

【０２２２】

１つの可能な実現形態において、トランザクション割り当てモードはハイブリッドのビジー度に基づく割り当てを含み、第１決定ユニット５０１はさらに、第１ノード機器のトランザクション処理数、第１ノード機器の機器リソース使用率、トランザクション処理数の重み、機器リソース使用率の重み、及び重み調節パラメータに基づいて、第１ノード機器に対応するトランザクション割り当て指標を決定することに用いられ、第１ノード機器は少なくとも２つのノード機器のうちのいずれか１つのノード機器である。

【０２２３】

１つの可能な実現形態において、該装置は、

協調ノード機器の機器識別情報を割り当て要求を発した端末に送信することに用いられる送信ユニットであって、端末は協調ノード機器の機器識別情報に基づき、ターゲットトランザクションのトランザクション情報を協調ノード機器に送信し、協調ノード機器によってトランザクション情報に基づいてターゲットトランザクションを協調処理することに用いられる、送信ユニットをさらに含む。

【０２２４】

１つの可能な実現形態において、分散型データベースシステムはキー値型のデータ記憶

フォーマットとセグメントページ型のデータ記憶フォーマットをサポートする。

【０２２５】

本願の実施例において、各ノード機器のそれぞれに対応するトランザクション割り当て指標に基づきターゲットトランザクションを協調処理することに用いられる協調ノード機器を決定し、トランザクションの割り当て過程はトランザクションに関するデータ項目を考慮する必要がなく、データ項目の分布状況を考慮する必要もない。このような方式に基づいて、個々のノード機器はいずれも分散化された機器としてトランザクションを協調処理することを可能にすることにより、トランザクションをノード間で処理することを可能にし、トランザクションの処理効率を向上させることに有利であり、トランザクション処理の信頼性が比較的高く、データベースシステムのシステム性能を高めることに有利である。

【０２２６】

図６に参照されるように、本願の実施例はトランザクション処理装置を提供し、該装置は、

ターゲットトランザクションのトランザクション情報を取得することに用いられる取得ユニット６０１と、

ターゲットトランザクションのトランザクション情報に基づいて、データノード機器にデータ読み取り要求を送信することに用いられる第１送信ユニット６０２であって、データノード機器は同一の記憶システムを共有する少なくとも２つのノード機器のうちターゲットトランザクションの処理に関与することに用いられるノード機器である、第１送信ユニット６０２と、

データノード機器から返信されたデータ読み取り結果がトランザクション認証条件を満たすことに応答して、データノード機器にトランザクション認証要求、及びローカル書き込みセットを送信することに用いられる第２送信ユニット６０３と、

データノード機器から返信されたターゲットトランザクションの認証結果に基づいて、ターゲットトランザクションの処理命令を決定することに用いられる決定ユニット６０４と、

データノード機器に処理命令を送信することに用いられる第３送信ユニット６０５であって、処理命令はコミット命令、又はアボート命令であり、データノード機器は処理命令を実行することに用いられる、第３送信ユニット６０５と、を含む。

【０２２７】

１つの可能な実現形態において、データ読み取り結果は第２論理ライフサイクルを携えており、第２論理ライフサイクルはデータノード機器によってデータ読み取り要求が携えているターゲットトランザクションの第１論理ライフサイクルに基づき決定され、第１論理ライフサイクルはタイムスタンプ下限、及びタイムスタンプ上限で構成され、第２送信ユニット６０３は、第１論理ライフサイクルのタイムスタンプ下限と第２論理ライフサイクルのタイムスタンプ下限における最大値をターゲットトランザクションの第３論理ライフサイクルのタイムスタンプ下限として用いることと、第１論理ライフサイクルのタイムスタンプ上限と第２論理ライフサイクルのタイムスタンプ上限における最小値をターゲットトランザクションの第３論理ライフサイクルのタイムスタンプ上限として用いることと、第３論理ライフサイクルが有効であることに応答して、データノード機器に第３論理ライフサイクルを携えているトランザクション認証要求を送信することであって、第３論理ライフサイクルが有効であることは、第３論理ライフサイクルのタイムスタンプ下限が第３論理ライフサイクルのタイムスタンプ上限よりも小さいことを指示することに用いられる、ことと、に用いられる。

【０２２８】

１つの可能な実現形態において、データノード機器の数は少なくとも２つであり、決定ユニット６０４は、少なくとも２つのデータノード機器から返信された少なくとも２つの認証結果において認証に合格しなかったことを指示することに用いられる認証結果が存在することに応答して、アボート命令をターゲットトランザクションの処理命令として用い

ることと、少なくとも２つのデータノード機器から返信された少なくとも２つの認証結果はいずれも認証に合格したことを指示することに応答して、少なくとも２つの認証結果を携えている論理ライフサイクルの共通部分をターゲット論理ライフサイクルとして用いることと、ターゲット論理ライフサイクルが有効であることに応答して、コミット命令をターゲットトランザクションの処理命令として用いることと、ターゲット論理ライフサイクルが無効であることに応答して、アボート命令をターゲットトランザクションの処理命令として用いることと、に用いられる。

【０２２９】

本願の実施例において、各ノード機器のそれぞれに対応するトランザクション割り当て指標に基づきターゲットトランザクションを協調処理することに用いられる協調ノード機器を決定し、トランザクションの割り当て過程はトランザクションに関するデータ項目を考慮する必要がなく、データ項目の分布状況を考慮する必要もない。このような方式に基づいて、個々のノード機器はいずれも分散化された機器としてトランザクションを協調処理することを可能にすることにより、トランザクションをノード間で処理することを可能にし、トランザクションの処理効率を向上させることに有利であり、トランザクション処理の信頼性が比較的高く、データベースシステムのシステム性能を高めることに有利である。

【０２３０】

図７に参照されるように、本願の実施例はトランザクション処理装置を提供し、該装置は、

協調ノード機器から送信されたデータ読み取り要求に基づいて、データ読み取り結果を取得することに用いられる第１取得ユニット７０１であって、協調ノード機器は同一の記憶システムを共有する少なくとも２つのノード機器のうちターゲットトランザクションを協調処理することに用いられるノード機器であり、協調ノード機器は少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づき決定される、第１取得ユニット７０１と、

データ読み取り結果を協調ノード機器に返信することに用いられる返信ユニット７０２と、

協調ノード機器から送信されたトランザクション認証要求、及びローカル書き込みセットに基づいて、ターゲットトランザクションの認証結果を取得することに用いられる第２取得ユニット７０３と、

協調ノード機器から送信されたターゲットトランザクションの処理命令を受信したことに応答して、処理命令を実行することに用いられる実行ユニット７０４であって、処理命令はコミット命令、又はアボート命令である、実行ユニット７０４と、を含み、

返信ユニット７０２はさらに、ターゲットトランザクションの認証結果を協調ノード機器に返信することに用いられる。

【０２３１】

１つの可能な実現形態において、データ読み取り要求はターゲットトランザクションの第１論理ライフサイクルを携えており、第１論理ライフサイクルはタイムスタンプ下限、及びタイムスタンプ上限で構成され、第１取得ユニット７０１は、第１論理ライフサイクルに基づいて、データ読み取り要求によって指示された読み取り対象のデータ項目の可視バージョンデータを決定することと、可視バージョンデータの作成タイムスタンプ、及び第１論理ライフサイクルに基づいて、ターゲットトランザクションの第２論理ライフサイクルを決定することと、第２論理ライフサイクル、及び可視バージョンデータを携えている結果をデータ読み取り結果として用いることと、に用いられる。

【０２３２】

１つの可能な実現形態において、トランザクション認証要求はターゲットトランザクションの第３論理ライフサイクルを携えており、第３論理ライフサイクルは協調ノード機器によって第１論理ライフサイクル、及び第２論理ライフサイクルに基づいて決定された有効論理ライフサイクルであり、第２取得ユニット７０３は、第３論理ライフサイクルのタ

イムスタンプ下限と第２論理ライフサイクルのタイムスタンプ下限における最大値をターゲットトランザクションの第４論理ライフサイクルのタイムスタンプ下限として用いること、第３論理ライフサイクルのタイムスタンプ上限と第２論理ライフサイクルのタイムスタンプ上限における最小値をターゲットトランザクションの第４論理ライフサイクルのタイムスタンプ上限として用いることと、第４論理ライフサイクルが有効であることに応答して、ローカル書き込みセットに対応する各書き込み対象のデータ項目の読み取りトランザクション関連情報及び第４論理ライフサイクルに基づいて、ターゲットトランザクションの第５論理ライフサイクルを決定することと、第５論理ライフサイクルが有効であることに応答して、認証に合格したことを指示することに用いられる認証結果をターゲットトランザクションの認証結果として用いることと、第５論理ライフサイクルが無効であることに応答して、認証に合格しなかったことを指示することに用いられる認証結果をターゲットトランザクションの認証結果として用いることと、に用いられる。

【０２３３】

１つの可能な実現形態において、１つの書き込み対象のデータ項目の読み取りトランザクション関連情報は該１つの書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプを含み、該１つの書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプは、該１つの書き込み対象のデータ項目を読み取った各読み取りトランザクションの論理コミットタイムスタンプにおける最大値を指示することに用いられ、第２取得ユニット７０３はさらに、各書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプ、及び第４論理ライフサイクルに基づいて、ターゲットトランザクションの第５論理ライフサイクルを決定することに用いられ、第５論理ライフサイクルのタイムスタンプ下限は各書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプにおける最大値よりも大きい。

【０２３４】

１つの可能な実現形態において、１つの書き込み対象のデータ項目の読み取りトランザクション関連情報は該１つの書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプを含み、ターゲット読み取りトランザクションはローカル認証に合格するか、又はコミット段階にある読み取りトランザクションであり、ターゲット読み取りトランザクションの終了タイムスタンプはターゲット読み取りトランザクションの論理ライフサイクルのタイムスタンプ上限であり、第２取得ユニット７０３はさらに、各書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプ、及び第４論理ライフサイクルに基づいて、ターゲットトランザクションの第５論理ライフサイクルを決定することに用いられ、第５論理ライフサイクルのタイムスタンプ下限は各書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプにおける最大値よりも大きい。

【０２３５】

本願の実施例において、各ノード機器のそれぞれに対応するトランザクション割り当て指標に基づきターゲットトランザクションを協調処理することに用いられる協調ノード機器を決定し、トランザクションの割り当て過程はトランザクションに関するデータ項目を考慮する必要がなく、データ項目の分布状況を考慮する必要もない。このような方式に基づいて、個々のノード機器はいずれも分散化された機器としてトランザクションを協調処理することを可能にすることにより、トランザクションをノード間で処理することを可能にし、トランザクションの処理効率を向上させることに有利であり、トランザクション処理の信頼性が比較的高く、データベースシステムのシステム性能を高めることに有利である。

【０２３６】

説明する必要があるように、上記実施例が提供する装置はその機能を実現する場合、上記各機能ユニットの分割のみを例として説明したが、実際の応用において、必要に基づき上記機能を異なる機能ユニットに割り当てて完了させることができ、すなわち機器の内部構造を異なる機能ユニットに分割して、以上記述された全部又は一部の機能を完了させる

。また、上記実施例が提供する装置は方法実施例と同一の構想に属し、その具体的な実現過程について、詳しくは方法実施例を参照すればよく、ここでは詳細な説明を省略する。

【０２３７】

図８は本願の実施例が提供するコンピュータ機器の構造模式図であり、該コンピュータ機器は配置又は性能が異なるため比較的大きく異なってもよく、１つ又は複数のプロセッサ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔｓ、ＣＰＵ）８０１と、１つ又は複数のメモリ８０２とを含んでもよく、ここで、該１つ又は複数のメモリ８０２において少なくとも１つのコンピュータプログラムが記憶され、該少なくとも１つのコンピュータプログラムは該１つ又は複数のプロセッサ８０１によってロードされて実行され、コンピュータ機器に上記各方法実施例が提供するトランザクション処理方法を実現させる。もちろん、該コンピュータ機器はさらに、入出力を行うように、有線又は無線ネットワークインタフェース、キーボード及び入出力インタフェースなどの部材を有してもよく、該コンピュータ機器はさらに機器の機能を実現することに用いられる他の部材を含んでもよく、ここでは詳細な説明を省略する。

【０２３８】

例示的な実施例において、非一時的コンピュータ可読記憶媒体をさらに提供し、該非一時的コンピュータ可読記憶媒体において少なくとも１つのコンピュータプログラムが記憶され、該少なくとも１つのコンピュータプログラムはコンピュータ機器のプロセッサによってロードされて実行され、コンピュータに上記いずれか１つのトランザクション処理方法を実現させる。

【０２３９】

１つの可能な実現形態において、上記非一時的コンピュータ可読記憶媒体は読み取り専用メモリ（Ｒｅａｄ－Ｏｎｌｙ　Ｍｅｍｏｒｙ、ＲＯＭ）、ランダムアクセスメモリ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ、ＲＡＭ）、読み取り専用光ディスク（Ｃｏｍｐａｃｔ　Ｄｉｓｃ　Ｒｅａｄ－Ｏｎｌｙ　Ｍｅｍｏｒｙ、ＣＤ－ＲＯＭ）、磁気テープ、フロッピーディスク及び光データ記憶機器などであってもよい。

【０２４０】

例示的な実施例において、コンピュータプログラム製品、又はコンピュータプログラムをさらに提供し、該コンピュータプログラム製品、又はコンピュータプログラムはコンピュータ命令を含み、該コンピュータ命令はコンピュータ可読記憶媒体に記憶される。コンピュータ機器のプロセッサはコンピュータ可読記憶媒体から該コンピュータ命令を読み取り、プロセッサは該コンピュータ命令を実行することにより、該コンピュータ機器が上記いずれか１つのトランザクション処理方法を実行する。

【０２４１】

理解すべきであるように、本願における用語「少なくとも１つ」とは、１つ又は複数を指し、「複数」又は「少なくとも２つ」の意味はいずれも２つ又は２つ以上を指す。「及び／又は」は、関連対象の関連関係を記述するためのものであり、３種の関係が存在することを表すことができ、例えば、Ａ及び／又はＢは、Ａが単独で存在する、ＡとＢが同時に存在する、Ｂが単独で存在するという３種の状況を表すことができる。文字「／」は、一般的に前後の関連オブジェクトが「又は」の関係であることを表す。

【０２４２】

以上は本願の例示的な実施例に過ぎず、本願を制限することに用いられるものではなく、本願の精神及び原則内で行われた任意の修正、等価置換、改良などは、いずれも本願の保護範囲内に含まれるべきである。

【符号の説明】

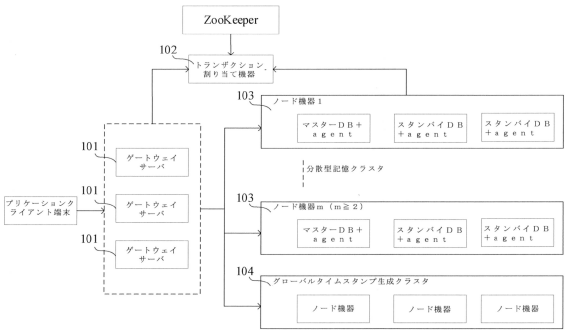
【０２４３】

１０１　ゲートウェイサーバ
１０２　機器
１０３　分散型記憶クラスタ
１０４　グローバルタイムスタンプ生成クラスタ

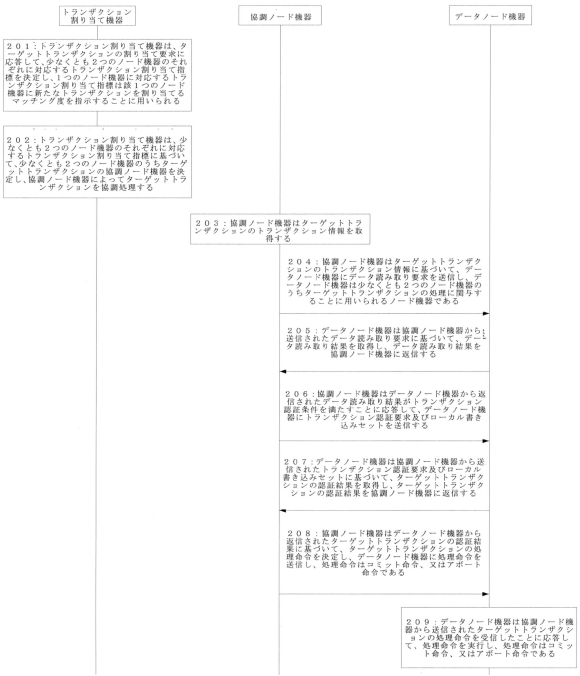
５０１　第１決定ユニット
５０２　第２決定ユニット
６０１　取得ユニット
６０２　第１送信ユニット
６０３　第２送信ユニット
６０４　決定ユニット
６０５　第３送信ユニット
７０１　第１取得ユニット
７０２　返信ユニット
７０３　第２取得ユニット
７０４　実行ユニット
８０１　プロセッサ
８０２　メモリ

【図面】

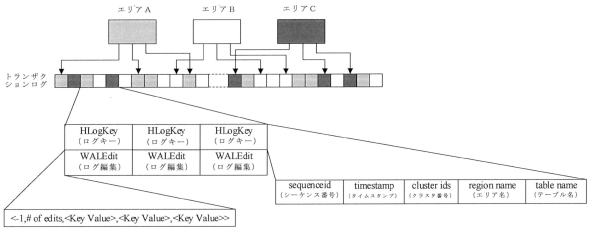
【図１】



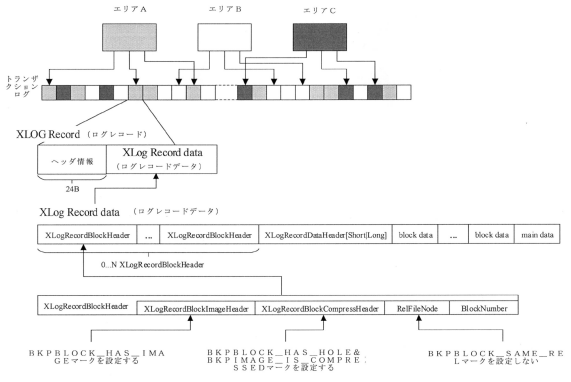
【図２】



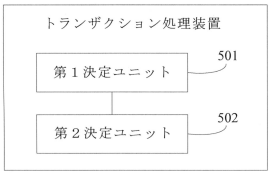
【図３】



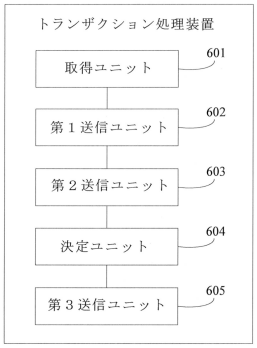
【図４】



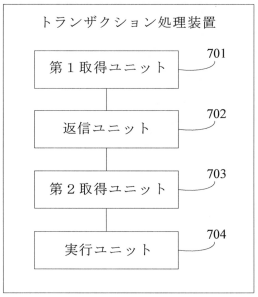
【図５】



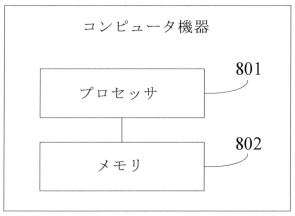
【図６】



【図７】



【図８】



【手続補正書】
【提出日】令和5年3月15日(2023.3.15)
【手続補正２】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
トランザクション割り当て機器が実行するトランザクション処理方法であって、前記トランザクション割り当て機器は分散型データベースシステム中にあり、前記分散型データベースシステムにおいて同一の記憶システムを共有する少なくとも２つのノード機器がさらに含まれ、前記方法は、
ターゲットトランザクションの割り当て要求に応答して、前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定するステップであって、１つのノード機器に対応するトランザクション割り当て指標は前記１つのノード機器に新たなトランザクションを割り当てるマッチング度を指示することに用いられる、ステップと、
前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づいて、前記少なくとも２つのノード機器のうち前記ターゲットトランザクションの協調ノード機器を決定し、前記協調ノード機器によって前記ターゲットトランザクションを協調処理するステップと、を含む、トランザクション処理方法。
【請求項２】
前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定する前記ステップは、
トランザクション割り当てモードを決定するステップであって、前記トランザクション割り当てモードはトランザクションのビジー度に基づく割り当て、機器のビジー度に基づく割り当て、及びハイブリッドのビジー度に基づく割り当てのうちのいずれか１つを含む、ステップと、
前記トランザクション割り当てモードによって指示された決定方式に基づき、前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定するステップと、を含む、請求項１に記載の方法。
【請求項３】
前記トランザクション割り当てモードはハイブリッドのビジー度に基づく割り当てを含み、前記トランザクション割り当てモードによって指示された決定方式に基づき、前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定する前記ステップは、
第１ノード機器のトランザクション処理数、前記第１ノード機器の機器リソース使用率、トランザクション処理数の重み、機器リソース使用率の重み、及び重み調節パラメータに基づいて、前記第１ノード機器に対応するトランザクション割り当て指標を決定するステップであって、前記第１ノード機器は前記少なくとも２つのノード機器のうちのいずれか１つのノード機器である、ステップを含む、請求項２に記載の方法。
【請求項４】
前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づいて、前記少なくとも２つのノード機器のうち前記ターゲットトランザクションの協調ノード機器を決定し、前記協調ノード機器によって前記ターゲットトランザクションを協調処理する前記ステップの後に、前記方法は、
前記協調ノード機器の機器識別情報を前記割り当て要求を発した端末に送信するステップであって、前記端末は前記協調ノード機器の機器識別情報に基づき、前記ターゲットトランザクションのトランザクション情報を前記協調ノード機器に送信し、前記協調ノード

機器によって前記トランザクション情報に基づいて前記ターゲットトランザクションを協調処理することに用いられる、ステップをさらに含む、請求項１～３のいずれか一項に記載の方法。

【請求項５】

前記分散型データベースシステムはキー値型のデータ記憶フォーマットとセグメントページ型のデータ記憶フォーマットをサポートする、請求項１～３のいずれか一項に記載の方法。

【請求項６】

協調ノード機器が実行するトランザクション処理方法であって、前記協調ノード機器は同一の記憶システムを共有する少なくとも２つのノード機器のうちターゲットトランザクションを協調処理することに用いられるノード機器であり、前記協調ノード機器は前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づき決定され、前記方法は、

前記ターゲットトランザクションのトランザクション情報を取得するステップと、

前記ターゲットトランザクションのトランザクション情報に基づいて、データノード機器にデータ読み取り要求を送信するステップであって、前記データノード機器は前記少なくとも２つのノード機器のうち前記ターゲットトランザクションの処理に関与することに用いられるノード機器である、ステップと、

前記データノード機器から返信されたデータ読み取り結果がトランザクション認証条件を満たすことに応答して、前記データノード機器にトランザクション認証要求、及びローカル書き込みセットを送信するステップと、

前記データノード機器から返信された前記ターゲットトランザクションの認証結果に基づいて、前記ターゲットトランザクションの処理命令を決定し、前記データノード機器に前記処理命令を送信するステップであって、前記処理命令はコミット命令、又はアボート命令であり、前記データノード機器は前記処理命令を実行することに用いられる、ステップと、を含む、トランザクション処理方法。

【請求項７】

前記データ読み取り結果は第２論理ライフサイクルを携えており、前記第２論理ライフサイクルは前記データノード機器によって前記データ読み取り要求が携えている前記ターゲットトランザクションの第１論理ライフサイクルに基づき決定され、前記第１論理ライフサイクルはタイムスタンプ下限、及びタイムスタンプ上限で構成され、

前記データノード機器から返信されたデータ読み取り結果がトランザクション認証条件を満たすことに応答して、前記データノード機器にトランザクション認証要求、及びローカル書き込みセットを送信する前記ステップは、

前記第１論理ライフサイクルのタイムスタンプ下限と前記第２論理ライフサイクルのタイムスタンプ下限における最大値を前記ターゲットトランザクションの第３論理ライフサイクルのタイムスタンプ下限として用い、前記第１論理ライフサイクルのタイムスタンプ上限と前記第２論理ライフサイクルのタイムスタンプ上限における最小値を前記ターゲットトランザクションの第３論理ライフサイクルのタイムスタンプ上限として用いるステップと、

前記第３論理ライフサイクルが有効であることに応答して、前記データノード機器に前記第３論理ライフサイクルを携えているトランザクション認証要求を送信するステップであって、前記第３論理ライフサイクルが有効であることは、前記第３論理ライフサイクルのタイムスタンプ下限が前記第３論理ライフサイクルのタイムスタンプ上限よりも小さいことを指示することに用いられる、ステップと、を含む、請求項６に記載の方法。

【請求項８】

前記データノード機器の数は少なくとも２つであり、前記データノード機器から返信された前記ターゲットトランザクションの認証結果に基づいて、前記ターゲットトランザクションの処理命令を決定し、前記データノード機器に前記処理命令を送信する前記ステップは、

前記少なくとも２つのデータノード機器から返信された少なくとも２つの認証結果において認証に合格しなかったことを指示することに用いられる認証結果が存在することに応答して、前記アボート命令を前記ターゲットトランザクションの処理命令として用いるステップと、

前記少なくとも２つのデータノード機器から返信された少なくとも２つの認証結果はいずれも認証に合格したことを指示することに応答して、前記少なくとも２つの認証結果が携えている論理ライフサイクルの共通部分をターゲット論理ライフサイクルとして用いるステップと、

前記ターゲット論理ライフサイクルが有効であることに応答して、前記コミット命令を前記ターゲットトランザクションの処理命令として用い、前記ターゲット論理ライフサイクルが無効であることに応答して、前記アボート命令を前記ターゲットトランザクションの処理命令として用いるステップと、を含む、請求項６又は７のいずれか一項に記載の方法。

【請求項９】

データノード機器が実行するトランザクション処理方法であって、前記データノード機器は同一の記憶システムを共有する少なくとも２つのノード機器のうちターゲットトランザクションの処理に関与することに用いられるノード機器であり、前記方法は、

協調ノード機器から送信されたデータ読み取り要求に基づいて、データ読み取り結果を取得し、前記データ読み取り結果を前記協調ノード機器に返信するステップであって、前記協調ノード機器は前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づき決定される、ステップと、

前記協調ノード機器から送信されたトランザクション認証要求、及びローカル書き込みセットに基づいて、前記ターゲットトランザクションの認証結果を取得し、前記ターゲットトランザクションの認証結果を前記協調ノード機器に返信するステップと、

前記協調ノード機器から送信された前記ターゲットトランザクションの処理命令を受信したことに応答して、前記処理命令を実行するステップであって、前記処理命令はコミット命令、又はアボート命令である、ステップと、を含む、トランザクション処理方法。

【請求項１０】

前記データ読み取り要求は前記ターゲットトランザクションの第１論理ライフサイクルを携えており、前記第１論理ライフサイクルはタイムスタンプ下限、及びタイムスタンプ上限で構成され、

協調ノード機器から送信されたデータ読み取り要求に基づいて、データ読み取り結果を取得し、前記データ読み取り結果を前記協調ノード機器に返信する前記ステップは、

前記第１論理ライフサイクルに基づいて、前記データ読み取り要求によって指示された読み取り対象のデータ項目の可視バージョンデータを決定するステップと、

前記可視バージョンデータの作成タイムスタンプ、及び前記第１論理ライフサイクルに基づいて、前記ターゲットトランザクションの第２論理ライフサイクルを決定するステップと、

前記第２論理ライフサイクル、及び前記可視バージョンデータが運ばれた結果を前記データ読み取り結果として用いるステップと、を含む、請求項９に記載の方法。

【請求項１１】

前記トランザクション認証要求は前記ターゲットトランザクションの第３論理ライフサイクルを携えており、前記第３論理ライフサイクルは前記協調ノード機器によって前記第１論理ライフサイクル、及び前記第２論理ライフサイクルに基づいて決定された有効論理ライフサイクルであり、

前記協調ノード機器から送信されたトランザクション認証要求、及びローカル書き込みセットに基づいて、前記ターゲットトランザクションの認証結果を取得する前記ステップは、

前記第３論理ライフサイクルのタイムスタンプ下限と前記第２論理ライフサイクルのタイムスタンプ下限における最大値を前記ターゲットトランザクションの第４論理ライフサ

イクルのタイムスタンプ下限として用い、前記第３論理ライフサイクルのタイムスタンプ上限と前記第２論理ライフサイクルのタイムスタンプ上限における最小値を前記ターゲットトランザクションの第４論理ライフサイクルのタイムスタンプ上限として用いるステップと、

前記第４論理ライフサイクルが有効であることに応答して、前記ローカル書き込みセットに対応する各書き込み対象のデータ項目の読み取りトランザクション関連情報及び前記第４論理ライフサイクルに基づいて、前記ターゲットトランザクションの第５論理ライフサイクルを決定するステップと、

前記第５論理ライフサイクルが有効であることに応答して、認証に合格したことを指示することに用いられる認証結果を前記ターゲットトランザクションの認証結果として用い、前記第５論理ライフサイクルが無効であることに応答して、認証に合格しなかったことを指示することに用いられる認証結果を前記ターゲットトランザクションの認証結果として用いるステップと、を含む、請求項１０に記載の方法。

【請求項１２】

１つの書き込み対象のデータ項目の読み取りトランザクション関連情報は前記１つの書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプを含み、前記１つの書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプは前記１つの書き込み対象のデータ項目を読み取った各読み取りトランザクションの論理コミットタイムスタンプにおける最大値を指示することに用いられ、

前記第４論理ライフサイクルが有効であることに応答して、前記ローカル書き込みセットに対応する各書き込み対象のデータ項目の読み取りトランザクション関連情報及び前記第４論理ライフサイクルに基づいて、前記ターゲットトランザクションの第５論理ライフサイクルを決定する前記ステップは、

前記各書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプ、及び前記第４論理ライフサイクルに基づいて、前記ターゲットトランザクションの第５論理ライフサイクルを決定するステップであって、前記第５論理ライフサイクルのタイムスタンプ下限は前記各書き込み対象のデータ項目の最大読み取りトランザクションタイムスタンプにおける最大値よりも大きい、ステップを含む、請求項１１に記載の方法。

【請求項１３】

１つの書き込み対象のデータ項目の読み取りトランザクション関連情報は前記１つの書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプを含み、前記ターゲット読み取りトランザクションはローカル認証に合格するか、又はコミット段階にある読み取りトランザクションであり、前記ターゲット読み取りトランザクションの終了タイムスタンプは前記ターゲット読み取りトランザクションの論理ライフサイクルのタイムスタンプ上限であり、

前記第４論理ライフサイクルが有効であることに応答して、前記ローカル書き込みセットに対応する各書き込み対象のデータ項目の読み取りトランザクション関連情報及び前記第４論理ライフサイクルに基づいて、前記ターゲットトランザクションの第５論理ライフサイクルを決定する前記ステップは、

前記各書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプ、及び前記第４論理ライフサイクルに基づいて、前記ターゲットトランザクションの第５論理ライフサイクルを決定するステップであって、前記第５論理ライフサイクルのタイムスタンプ下限は前記各書き込み対象のデータ項目のターゲット読み取りトランザクションの終了タイムスタンプにおける最大値よりも大きい、ステップを含む、請求項１１に記載の方法。

【請求項１４】

トランザクション処理システムであって、前記トランザクション処理システムは協調ノード機器と、データノード機器とを含み、前記協調ノード機器は同一の記憶システムを共有する少なくとも２つのノード機器のうちターゲットトランザクションを協調処理することに用いられるノード機器であり、前記協調ノード機器は前記少なくとも２つのノード機

器のそれぞれに対応するトランザクション割り当て指標に基づき決定され、前記データノード機器は前記少なくとも２つのノード機器のうち前記ターゲットトランザクションの処理に関与することに用いられるノード機器であり、
前記協調ノード機器は、前記ターゲットトランザクションのトランザクション情報を取得することと、前記ターゲットトランザクションのトランザクション情報に基づいて、前記データノード機器にデータ読み取り要求を送信することと、に用いられ、
前記データノード機器は、前記協調ノード機器から送信された前記データ読み取り要求に基づいて、データ読み取り結果を取得し、前記データ読み取り結果を前記協調ノード機器に返信することに用いられ、
前記協調ノード機器はさらに、前記データノード機器から返信された前記データ読み取り結果がトランザクション認証条件を満たすことに応答して、前記データノード機器にトランザクション認証要求、及びローカル書き込みセットを送信することに用いられ、
前記データノード機器はさらに、前記協調ノード機器から送信された前記トランザクション認証要求、及び前記ローカル書き込みセットに基づいて、前記ターゲットトランザクションの認証結果を取得し、前記ターゲットトランザクションの認証結果を前記協調ノード機器に返信することに用いられ、
前記協調ノード機器はさらに、前記データノード機器から返信された前記ターゲットトランザクションの認証結果に基づいて、前記ターゲットトランザクションの処理命令を決定し、前記データノード機器に前記処理命令を送信することに用いられ、前記処理命令はコミット命令、又はアボート命令であり、
前記データノード機器はさらに、前記協調ノード機器から送信された前記ターゲットトランザクションの処理命令を受信したことに応答して、前記処理命令を実行することに用いられる、トランザクション処理システム。
【請求項１５】
トランザクション処理装置であって、前記装置は、
ターゲットトランザクションの割り当て要求に応答して、同一の記憶システムを共有する少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標を決定することに用いられる第１決定ユニットであって、１つのノード機器に対応するトランザクション割り当て指標は前記１つのノード機器に新たなトランザクションを割り当てるマッチング度を指示することに用いられる、第１決定ユニットと、
前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づいて、前記少なくとも２つのノード機器のうち前記ターゲットトランザクションの協調ノード機器を決定し、前記協調ノード機器によって前記ターゲットトランザクションを協調処理することに用いられる第２決定ユニットと、を含む、トランザクション処理装置。
【請求項１６】
トランザクション処理装置であって、前記装置は、
ターゲットトランザクションのトランザクション情報を取得することに用いられる取得ユニットと、
前記ターゲットトランザクションのトランザクション情報に基づいて、データノード機器にデータ読み取り要求を送信することに用いられる第１送信ユニットであって、前記データノード機器は同一の記憶システムを共有する少なくとも２つのノード機器のうち前記ターゲットトランザクションの処理に関与することに用いられるノード機器である、第１送信ユニットと、
前記データノード機器から返信されたデータ読み取り結果がトランザクション認証条件を満たすことに応答して、前記データノード機器にトランザクション認証要求及び、ローカル書き込みセットを送信することに用いられる第２送信ユニットと、
前記データノード機器から返信された前記ターゲットトランザクションの認証結果に基づいて、前記ターゲットトランザクションの処理命令を決定することに用いられる決定ユニットと、

前記データノード機器に前記処理命令を送信することに用いられる第３送信ユニットであって、前記処理命令はコミット命令、又はアボート命令であり、前記データノード機器は前記処理命令を実行することに用いられる、第３送信ユニットと、を含む、トランザクション処理装置。

【請求項１７】

トランザクション処理装置であって、前記装置は、

協調ノード機器から送信されたデータ読み取り要求に基づいて、データ読み取り結果を取得することに用いられる第１取得ユニットであって、前記協調ノード機器は同一の記憶システムを共有する少なくとも２つのノード機器のうちターゲットトランザクションを協調処理することに用いられるノード機器であり、前記協調ノード機器は前記少なくとも２つのノード機器のそれぞれに対応するトランザクション割り当て指標に基づき決定される、第１取得ユニットと、

前記データ読み取り結果を前記協調ノード機器に返信することに用いられる返信ユニットと、

前記協調ノード機器から送信されたトランザクション認証要求、及びローカル書き込みセットに基づいて、前記ターゲットトランザクションの認証結果を取得することに用いられる第２取得ユニットと、

前記協調ノード機器から送信された前記ターゲットトランザクションの処理命令を受信したことに応答して、前記処理命令を実行することに用いられる実行ユニットであって、前記処理命令はコミット命令、又はアボート命令である、実行ユニットと、を含み、

前記返信ユニットはさらに、前記ターゲットトランザクションの認証結果を前記協調ノード機器に返信することに用いられる、トランザクション処理装置。

【請求項１８】

コンピュータ機器であって、前記コンピュータ機器はプロセッサと、メモリとを含み、前記メモリにおいて少なくとも１つのコンピュータプログラムが記憶され、前記少なくとも１つのコンピュータプログラムは前記プロセッサによってロードされて実行され、前記コンピュータ機器に、請求項１～５のいずれか一項に記載のトランザクション処理方法、又は請求項６～８のいずれか一項に記載のトランザクション処理方法、又は請求項９～１３のいずれか一項に記載のトランザクション処理方法を実現させる、コンピュータ機器。

【請求項１９】

コンピュータプログラムであって、プロセッサが前記コンピュータプログラムを実行することにより、コンピュータ機器が請求項１～５のいずれか一項に記載のトランザクション処理方法、請求項６～８のいずれか一項に記載のトランザクション処理方法、又は請求項９～１３のいずれか一項に記載のトランザクション処理方法を実行するように構成された、コンピュータプログラム。

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2021/126408

A. CLASSIFICATION OF SUBJECT MATTER

G06F 9/48(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, CNKI, WPI, EPODOC: 事务, 分配, 分发, 节点, 指标, 利用率, 使用率, 空闲, 验证, 本地写集, 回滚, transaction, assign, node, index, idle, verify, abort, utilization ratio

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 112162846 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 01 January 2021 (2021-01-01) claims 1-15	1-20
Y	CN 111597015 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD. et al.) 28 August 2020 (2020-08-28) description paragraphs 61-98, 140-144, 229-390	1-20
Y	CN 110287022 A (BEIJING DA MI TECHNOLOGY CO., LTD.) 27 September 2019 (2019-09-27) description paragraphs 2, 48-119	1-20
A	CN 108958942 A (ZHENGZHOU YUNHAI INFORMATION TECHNOLOGY CO., LTD.) 07 December 2018 (2018-12-07) entire document	1-20
A	US 2011153566 A1 (MICROSOFT CORPORATION) 23 June 2011 (2011-06-23) entire document	1-20

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
07 January 2022	**25 January 2022**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088, China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (January 2015)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2021/126408

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 112162846 A	01 January 2021	None	
CN 111597015 A	28 August 2020	None	
CN 110287022 A	27 September 2019	None	
CN 108958942 A	07 December 2018	None	
US 2011153566 A1	23 June 2011	US 8396831 B2	12 March 2013

Form PCT/ISA/210 (patent family annex) (January 2015)

国际检索报告

国际申请号
PCT/CN2021/126408

A. 主题的分类

G06F 9/48(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

CNPAT, CNKI, WPI, EPODOC: 事务，分配，分发，节点，指标，利用率，使用率，空闲，验证，本地写集，回滚，transaction, assign, node, index, idle, verify, abort, utilization ratio

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
PX	CN 112162846 A (腾讯科技深圳有限公司) 2021年1月1日 (2021-01-01) 权利要求1-15	1-20
Y	CN 111597015 A (腾讯科技深圳有限公司 等) 2020年8月28日 (2020-08-28) 说明书第61-98、140-144、229-390段	1-20
Y	CN 110287022 A (北京大米科技有限公司) 2019年9月27日 (2019-09-27) 说明书第2、48-119段	1-20
A	CN 108958942 A (郑州云海信息技术有限公司) 2018年12月7日 (2018-12-07) 全文	1-20
A	US 2011153566 A1 (MICROSOFT CORPORATION) 2011年6月23日 (2011-06-23) 全文	1-20

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：
"A" 认为不特别相关的表示了现有技术一般状态的文件
"E" 在国际申请日的当天或之后公布的在先申请或专利
"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）
"O" 涉及口头公开、使用、展览或其他方式公开的文件
"P" 公布日先于国际申请日但迟于所要求的优先权日的文件
"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件
"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性
"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性
"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2022年1月7日	2022年1月25日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 巩瑜 电话号码 86-(10)-53961382

PCT/ISA/210 表(第2页) (2015年1月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2021/126408

检索报告引用的专利文件	公布日（年/月/日）	同族专利	公布日（年/月/日）
CN 112162846 A	2021年1月1日	无	
CN 111597015 A	2020年8月28日	无	
CN 110287022 A	2019年9月27日	无	
CN 108958942 A	2018年12月7日	无	
US 2011153566 A1	2011年6月23日	US 8396831 B2	2013年3月12日

PCT/ISA/210 表(同族专利附件) (2015年1月)

フロントページの続き

MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,IT,JO,JP,KE,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,WS,ZA,ZM,ZW

（特許庁注：以下のものは登録商標）

１．ＵＮＩＸ