

# Cortex: Prometheus as a Service, One Year On

Tom Wilkie, PromCon 2017  
[tom.wilkie@gmail.com](mailto:tom.wilkie@gmail.com)



<https://github.com/weaveworks/cortex>







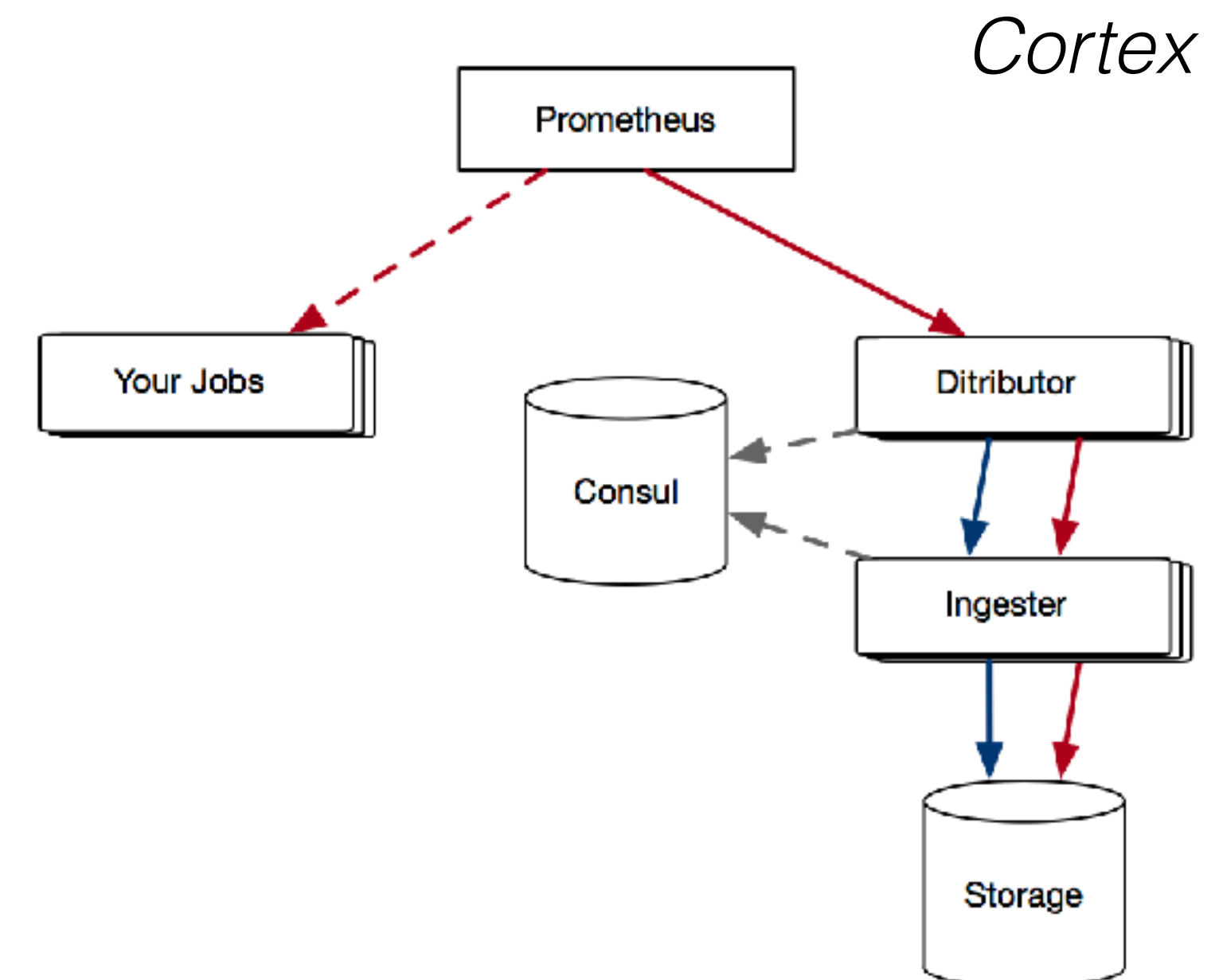
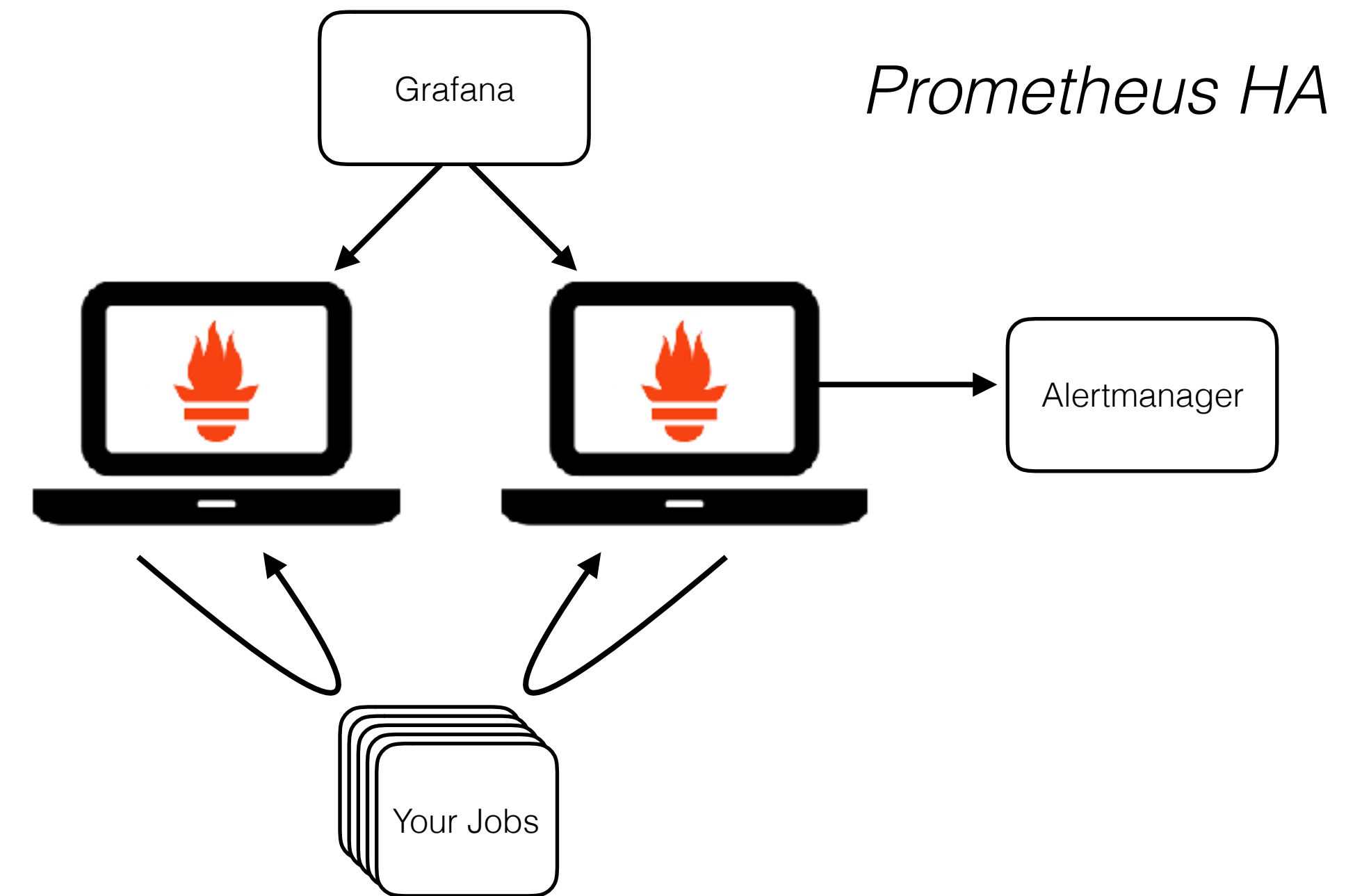
<https://www.youtube.com/watch?v=3Tb4Wc0kfCM>

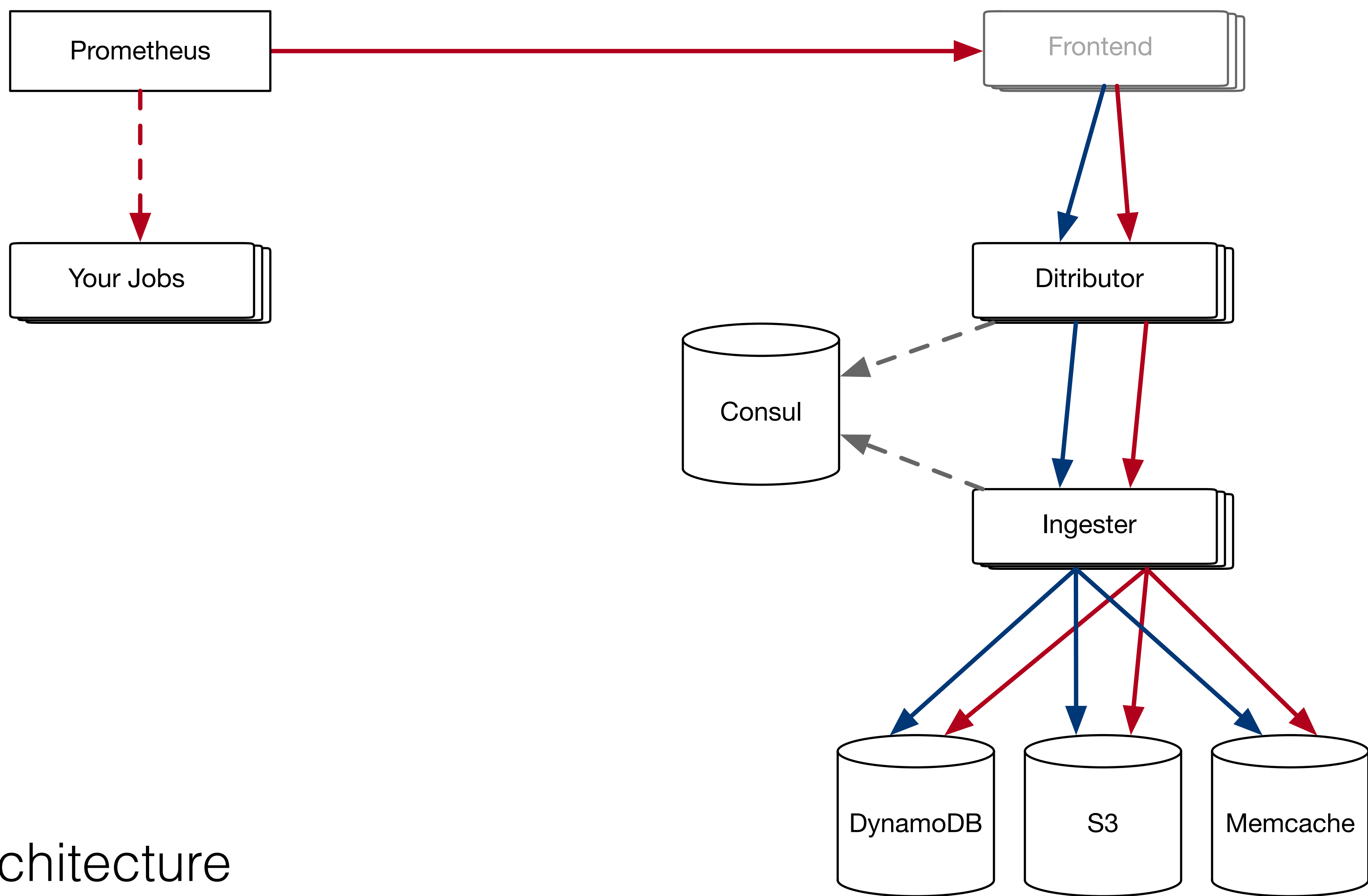
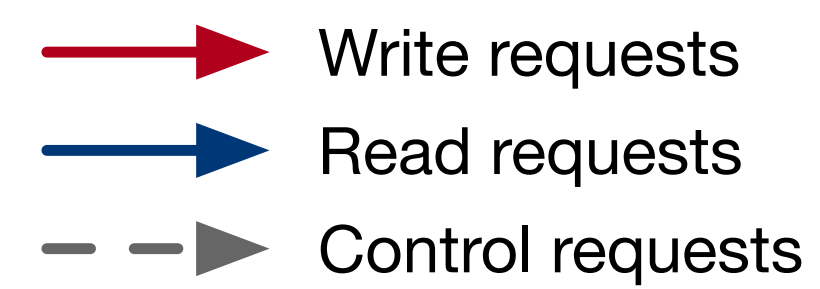
The screenshot shows a web browser window displaying a YouTube video. The browser's address bar shows the URL <https://www.youtube.com/watch?v=3Tb4Wc0kfCM>. The YouTube interface includes a search bar, navigation icons, and a user profile icon labeled 'Kausal'. The video content features a man in a light blue shirt speaking at a podium. Behind him is a presentation slide with the Prometheus logo (a red circle with a white flame) and the text 'PromCon 2016', 'd Sponsors', and 'Perception'. The video player controls at the bottom show a progress bar at 6:57 / 30:08, along with play, volume, and full-screen buttons. Below the video, the title 'PromCon 2016: Multitenant, Scale-Out Prometheus - Tom Wilkie' is displayed, along with the channel name 'Prometheus Monitoring' and a 'Subscribe' button with 722 subscribers. The video has 908 views and 3 likes. To the right, an 'Up next' section shows a video titled 'Tom Wilkie: Cortex: open-source, horizontally-scalable, distributed Prometheus' with 47 views. At the bottom of the browser window, another URL is visible: <https://www.youtube.com/watch?v=Xi4jq2IUbls>.



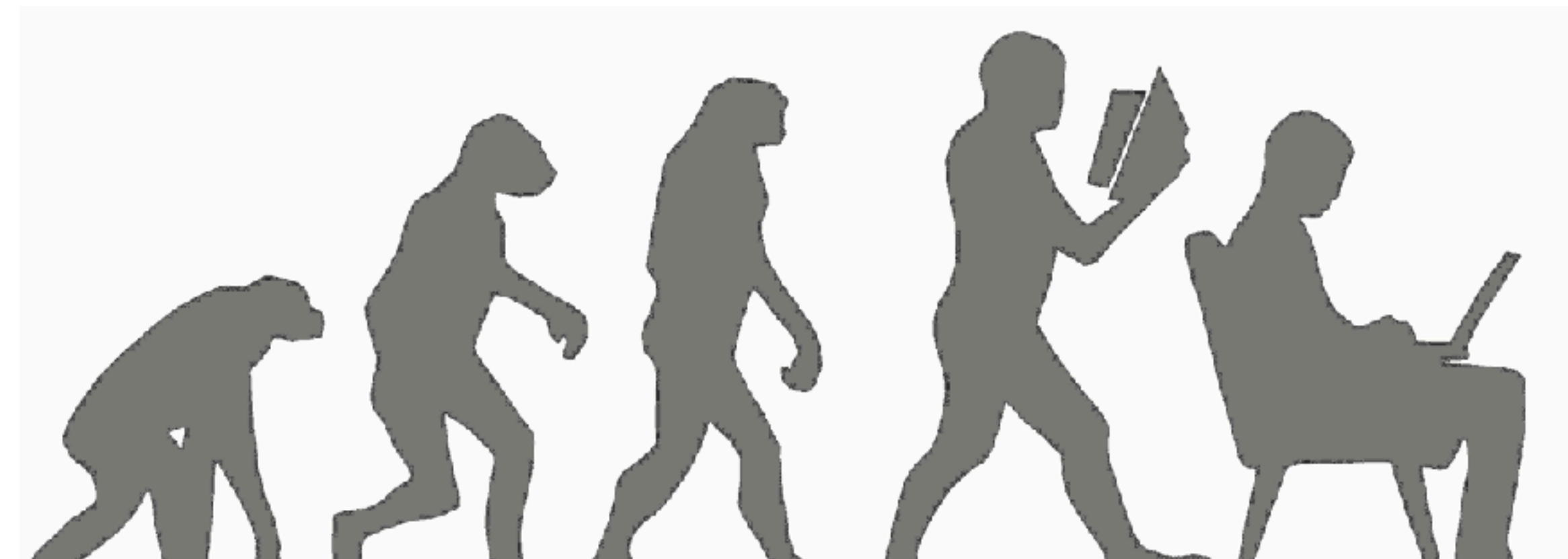
## Cortex: Prometheus as a Service

- Natively multi tenant; isolate different customers in the same services.
- Different story around scaling & HA
- “Virtually infinite” retention and durability
- Opportunities for performance enhancements





# A Year's Evolution

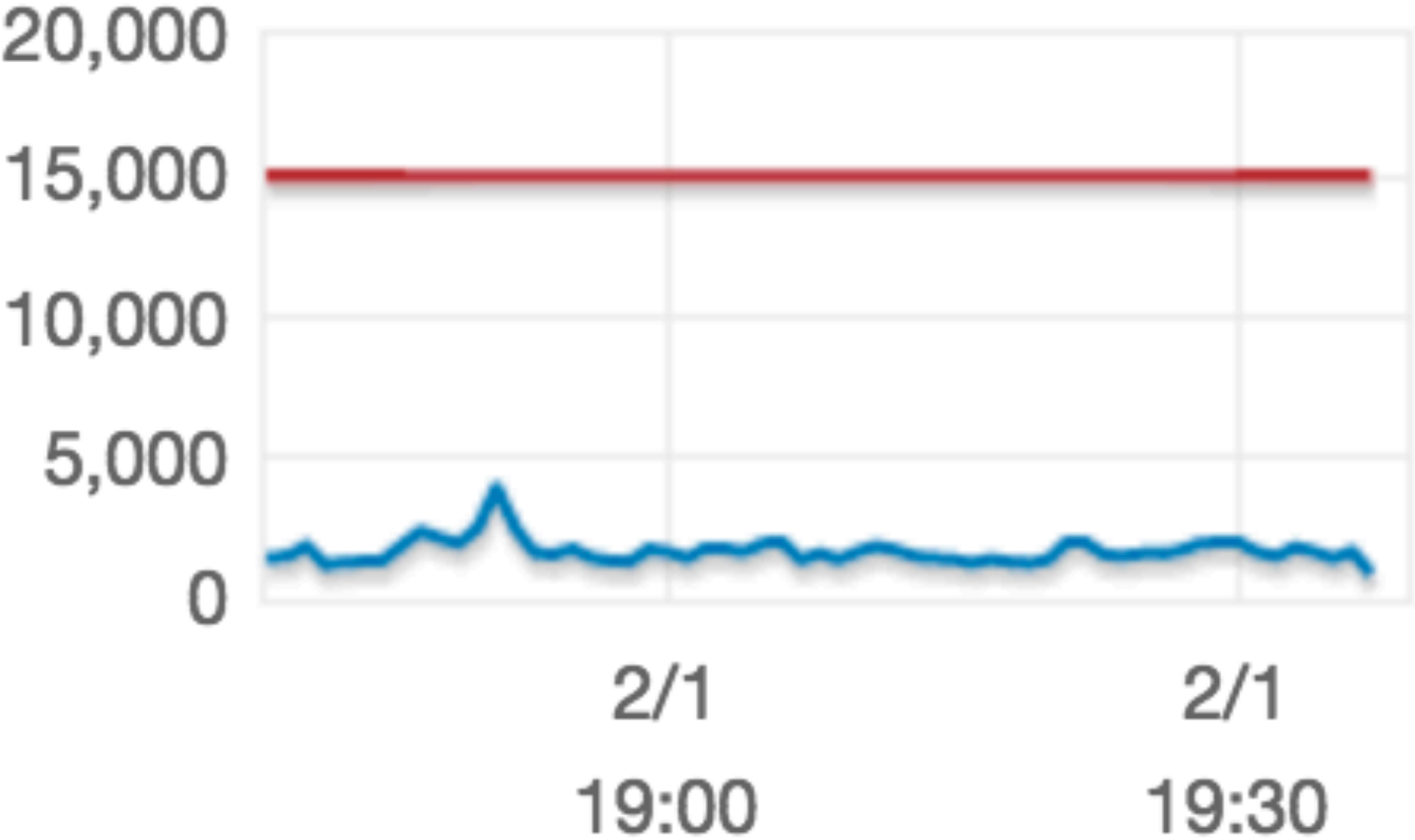


# Problem #1: DynamoDB Write Throughput



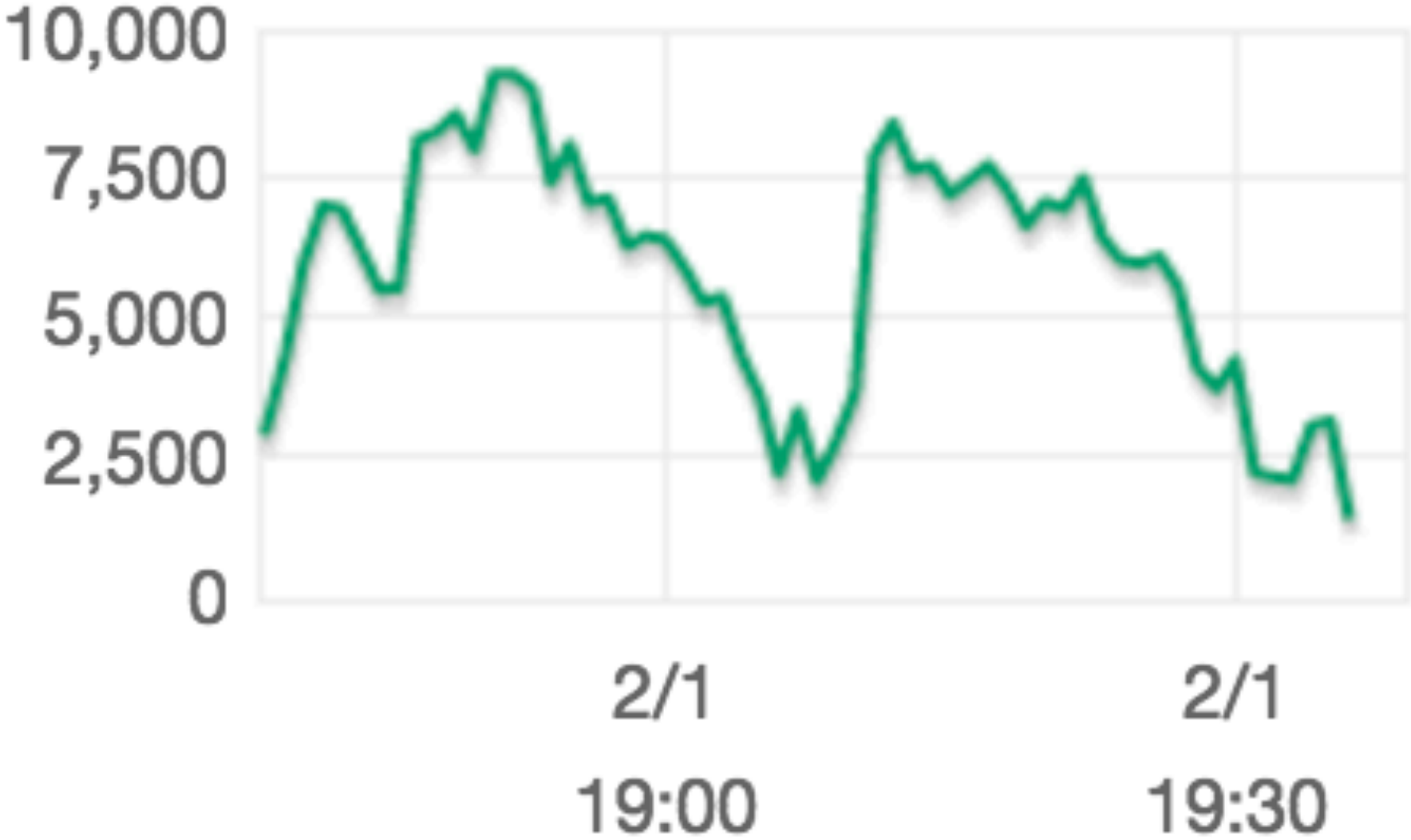


**Write capacity** (Units/Second - 1 min avg) 



 Provisioned  Consumed

**Throttled write requests** (Count)

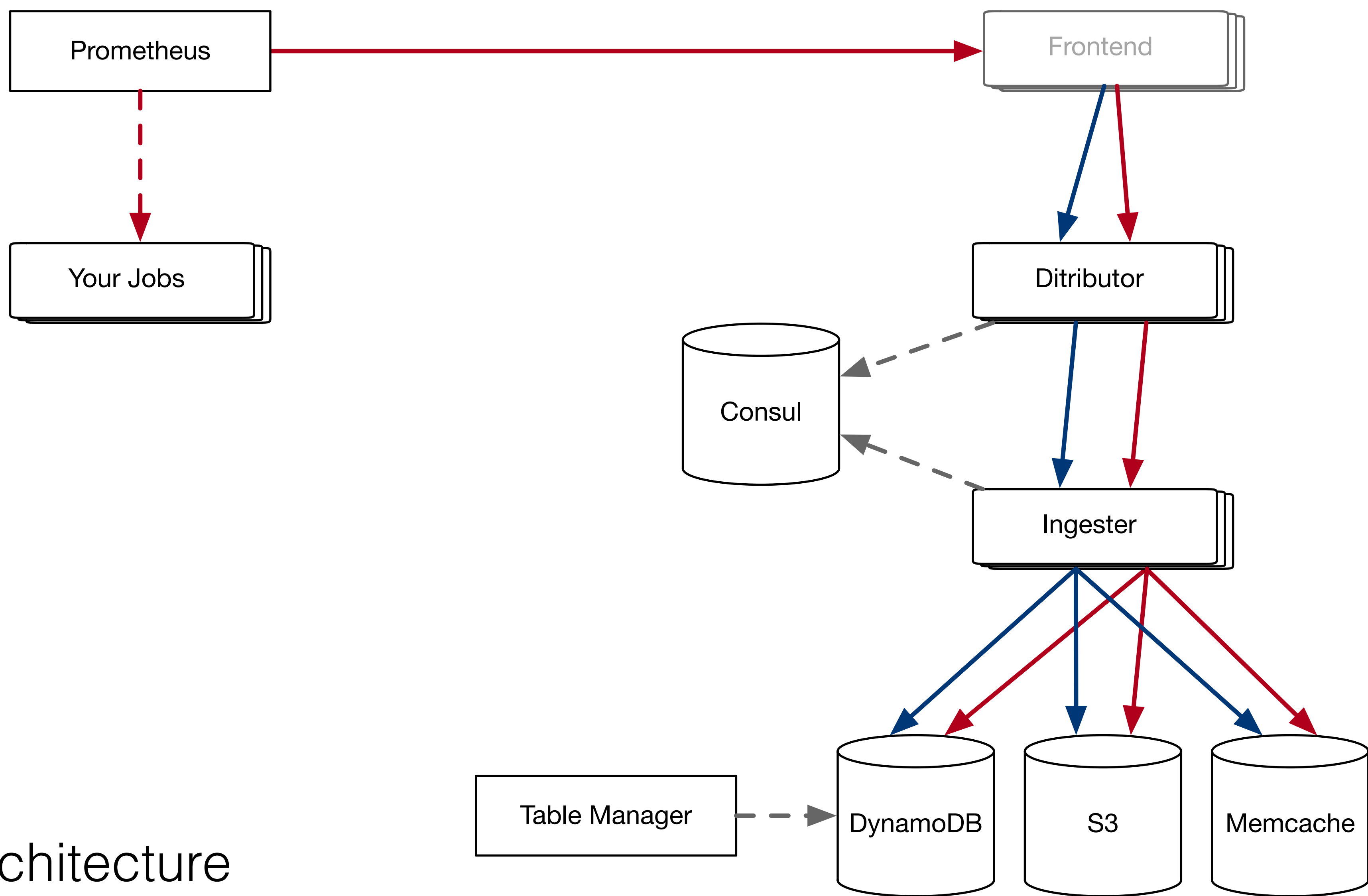
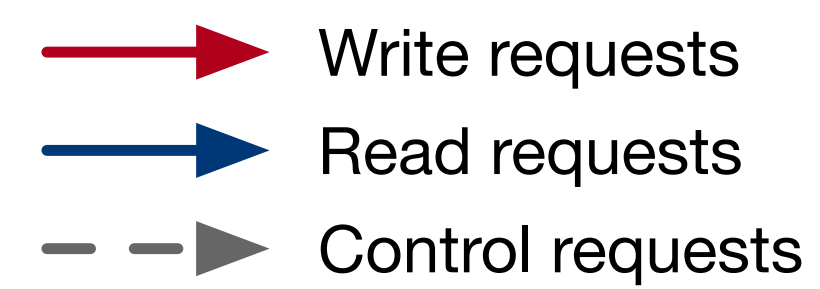


 Put  Update  Delete  Batch write

<https://github.com/weaveworks/cortex/issues/254>







# Problem #2: DynamoDB Write Throughput, again



Original schema:

- Hash Key: `<user ID>:<hour>:<metric name>`
- Range Key: `<label name>:<label value>:<chunk ID>`

New schema:

- Hash Key: `<user ID>:<day>:<metric name>:<label name>`
- Range Key: `<chunk ID>:<chunk end time>`

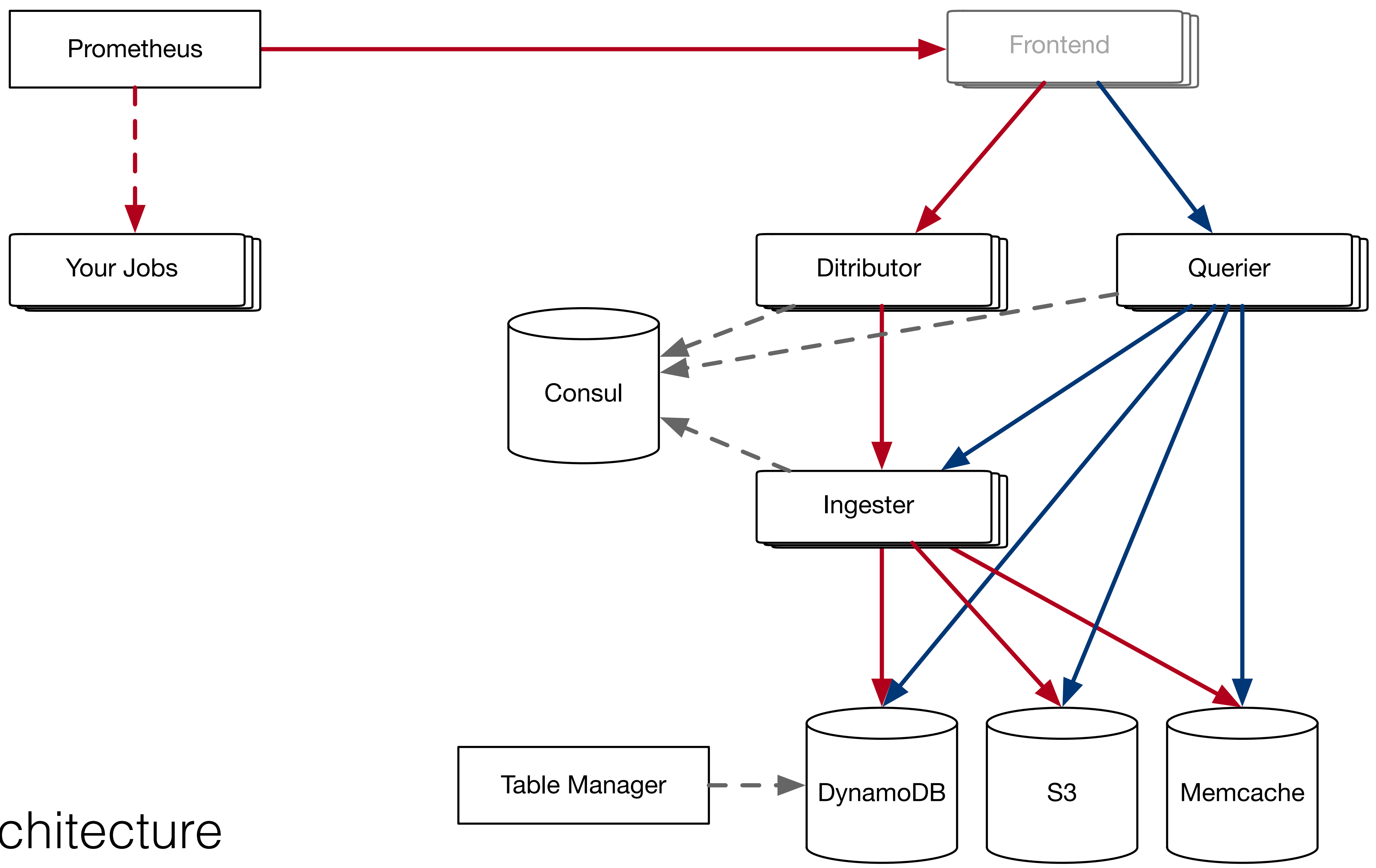
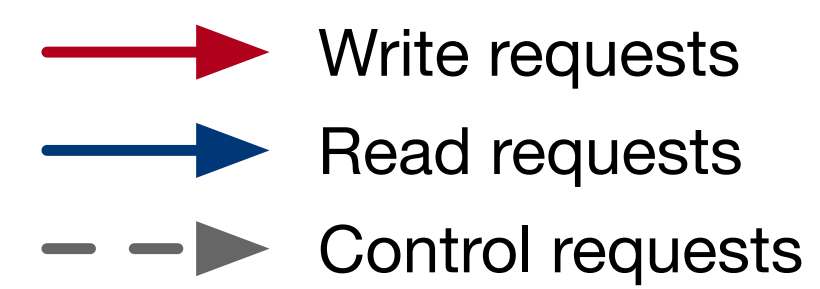


<https://github.com/weaveworks/cortex/pull/262>



# Problem #3: Queries of Death

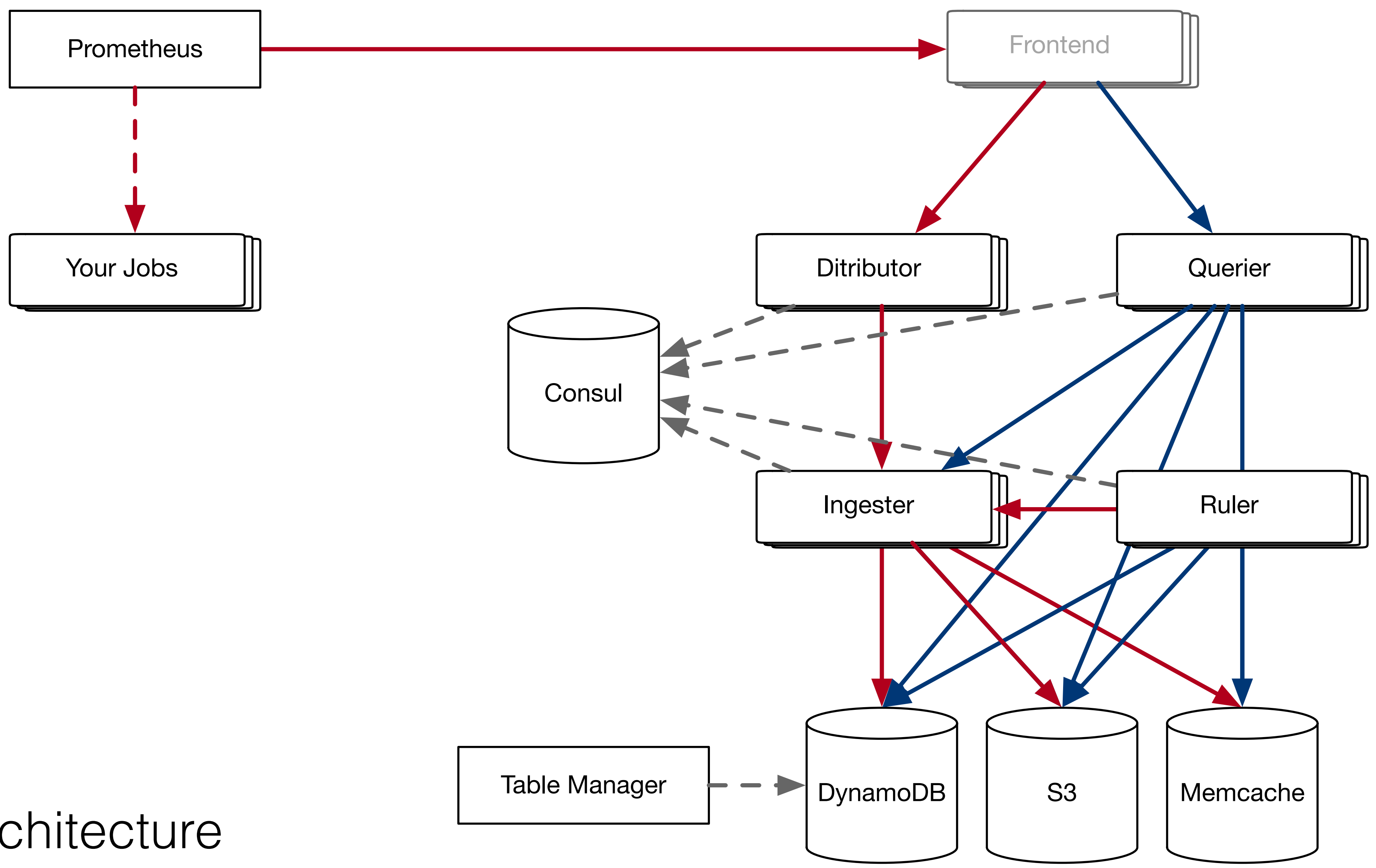
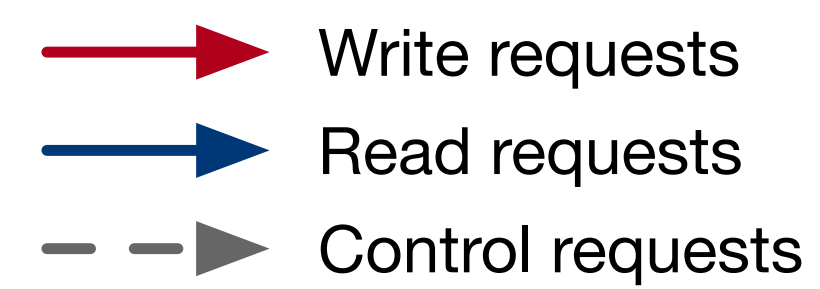




# Problem #3: Recording rules and alerts

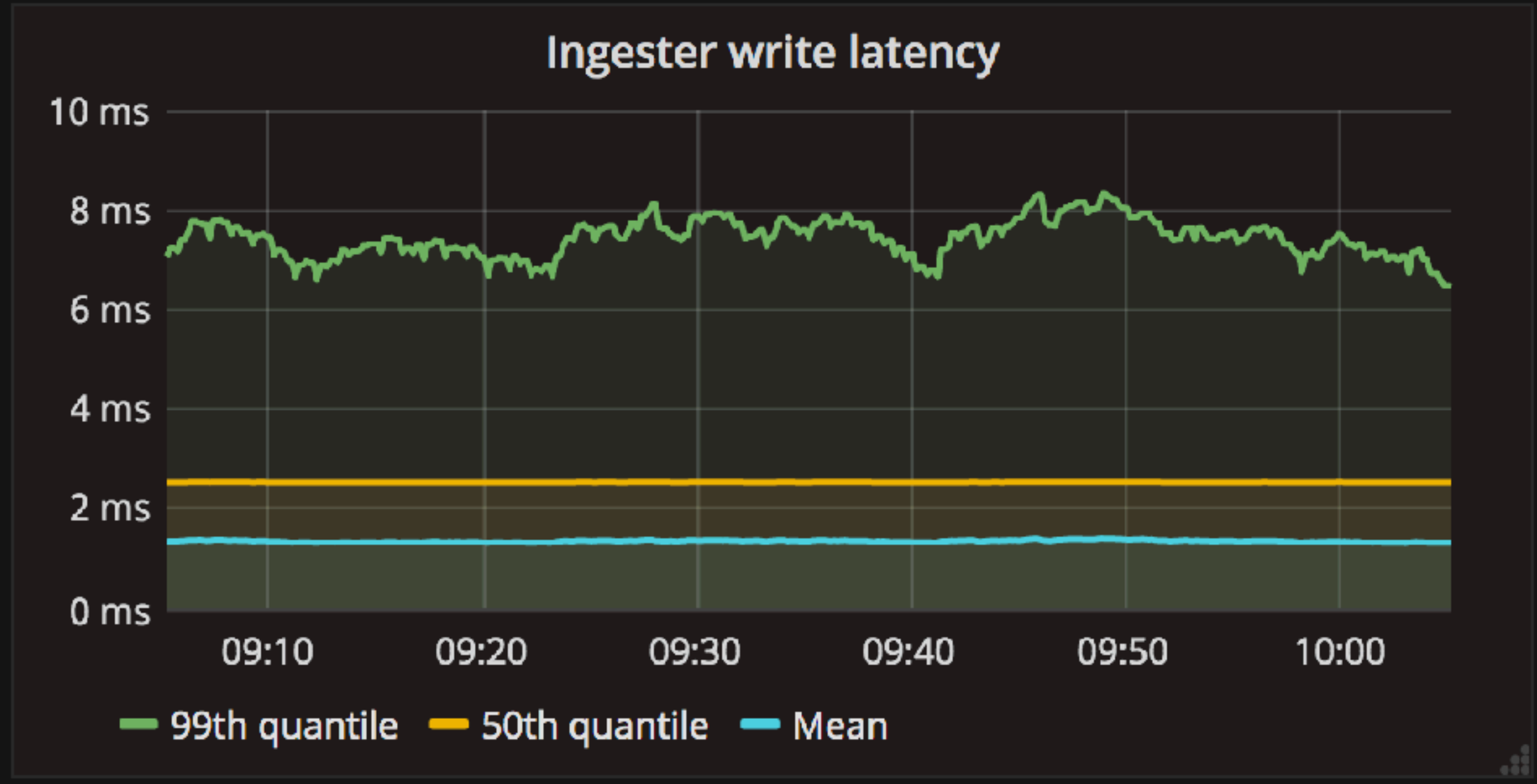
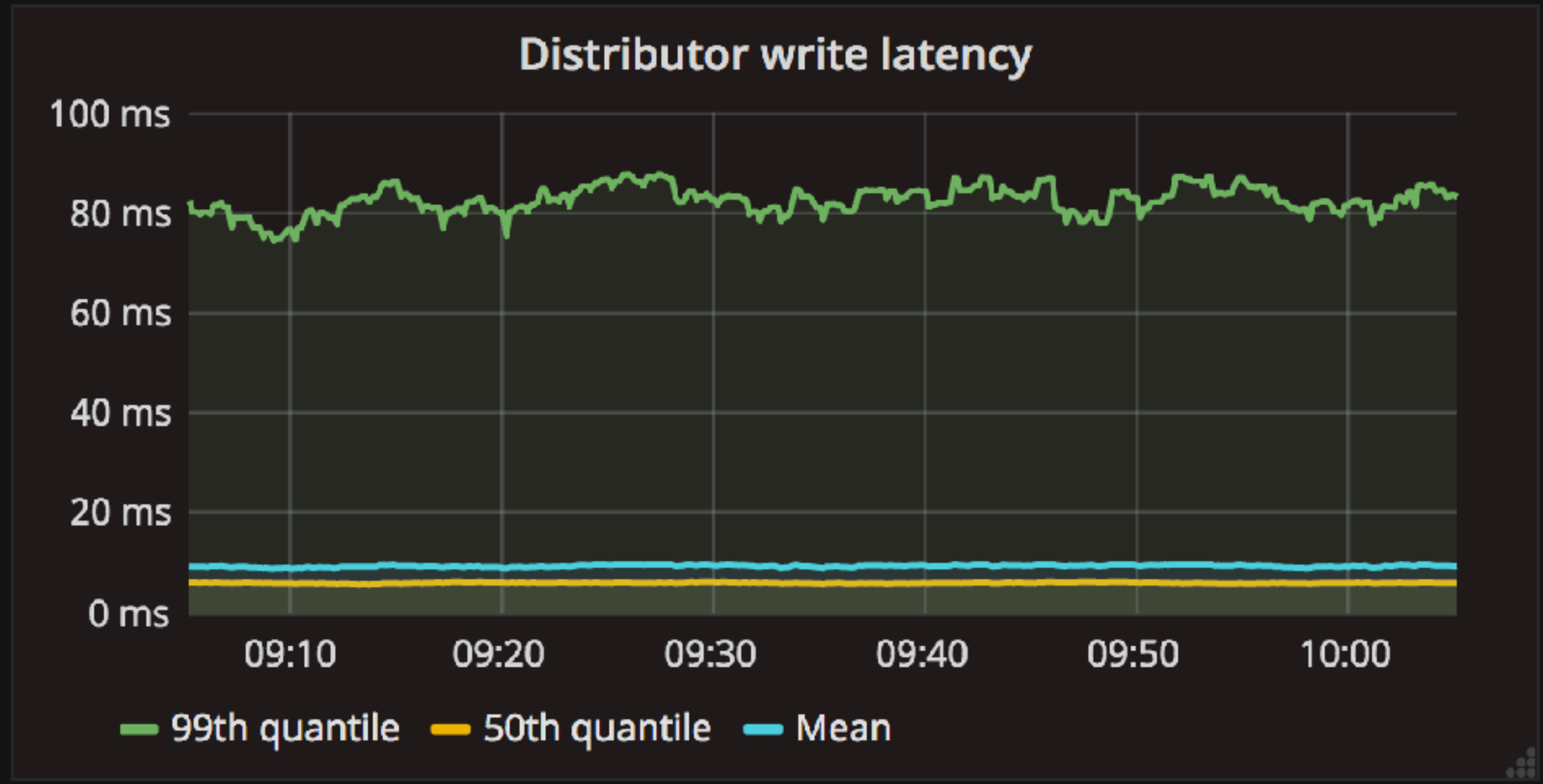






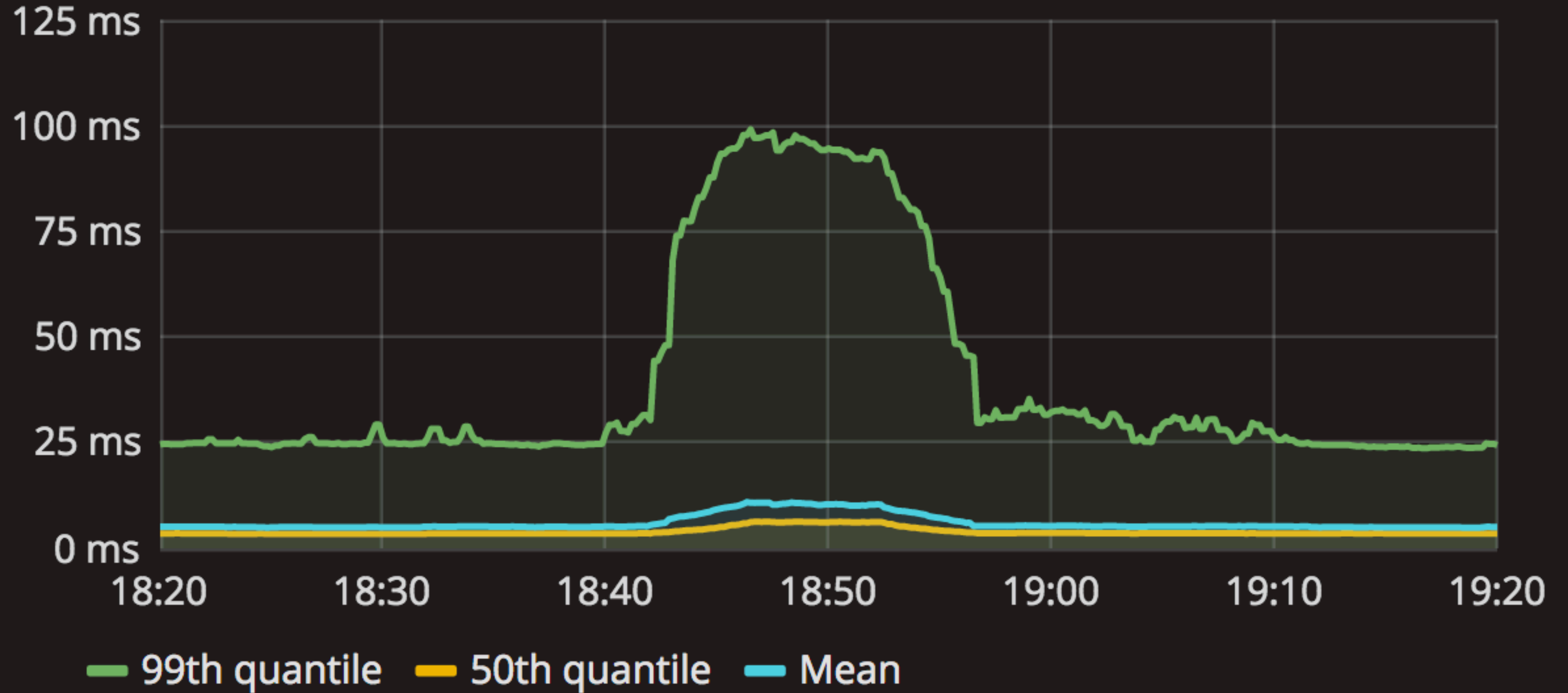
# Problem #4: Long tail







# Distributor write latency



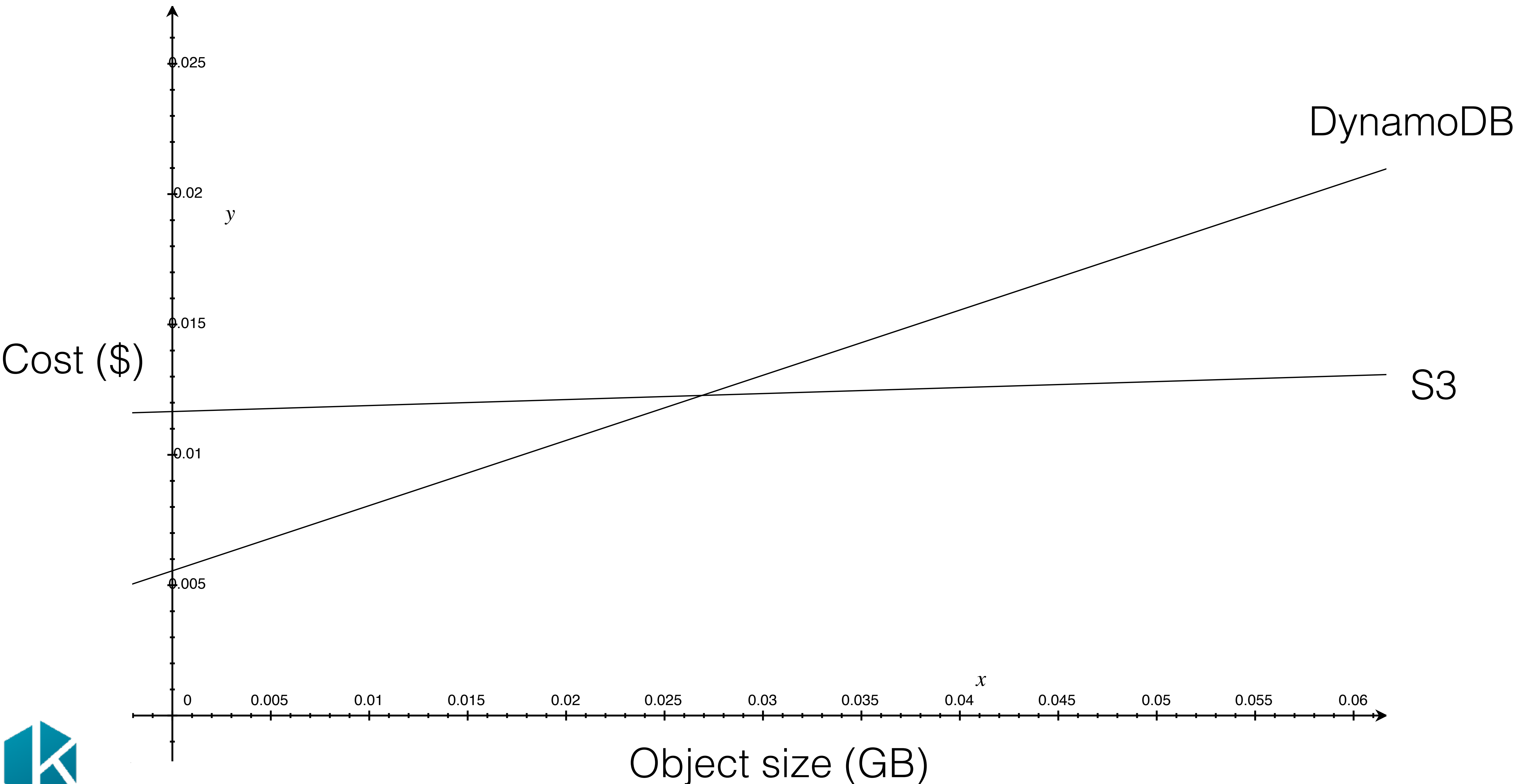
# Problem #5: Cost

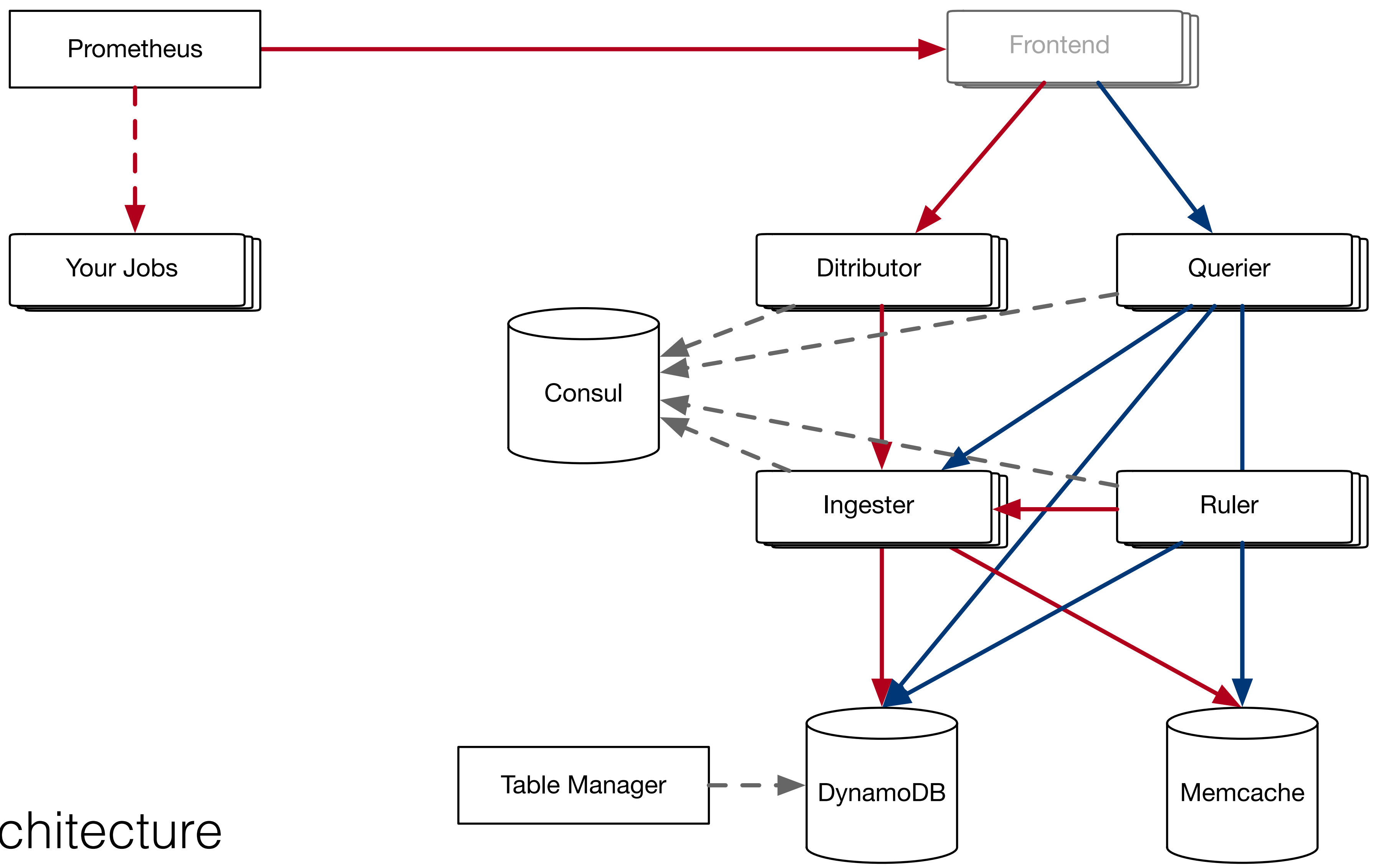
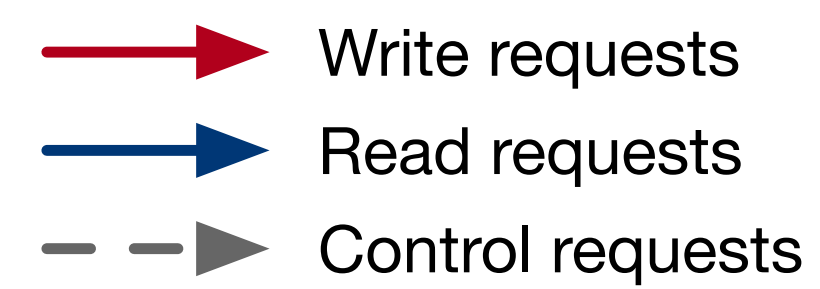


	<b>S3</b>	<b>DynamoDB</b>
<b>IOP Cost (\$/IOP)</b>	$5 \times 10^{-6}$	$2 \times 10^{-7}$
<b>Storage Cost (\$/GB/Month)</b>	0.023	0.250

<https://github.com/weaveworks/cortex/issues/141>



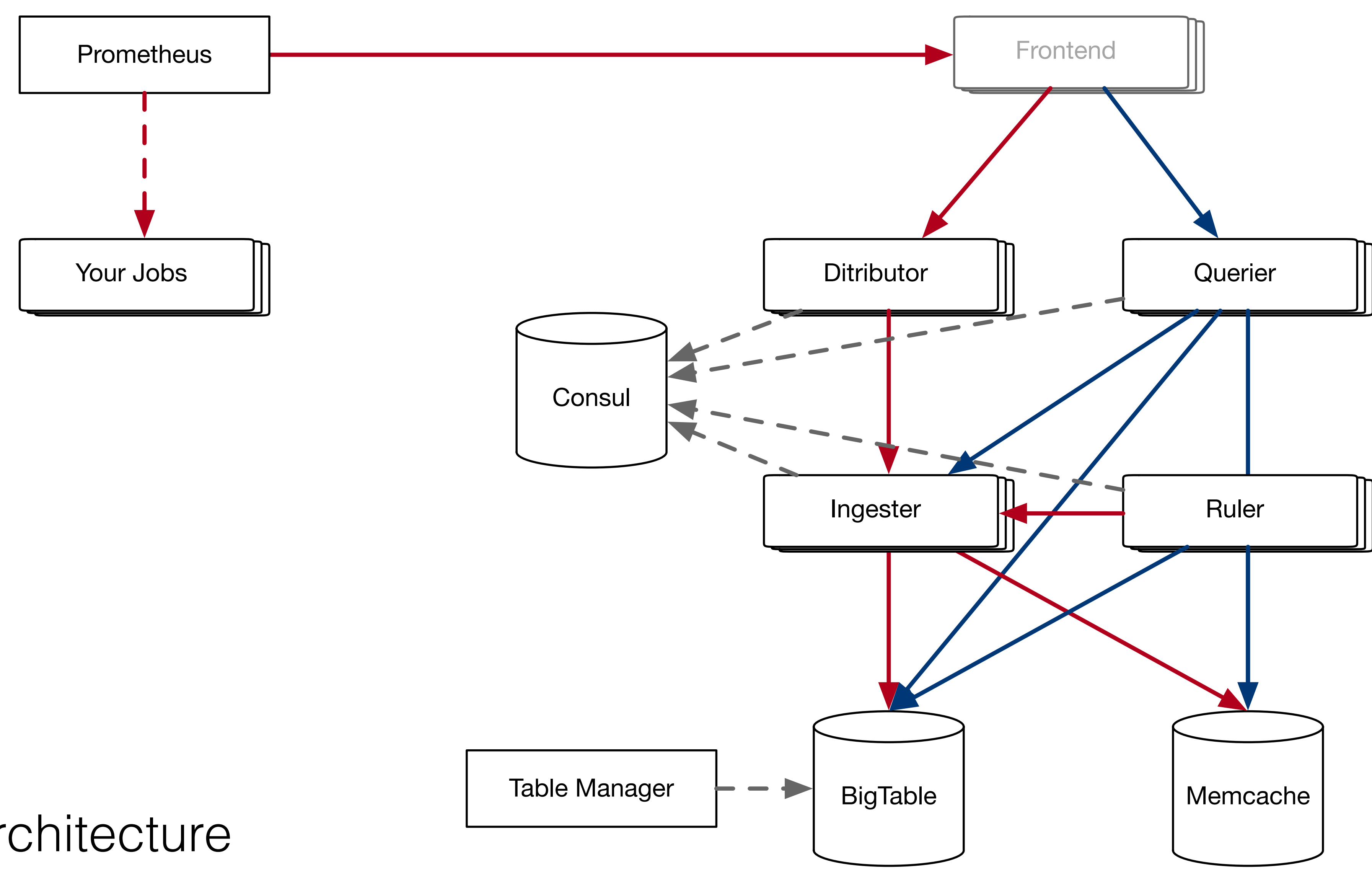
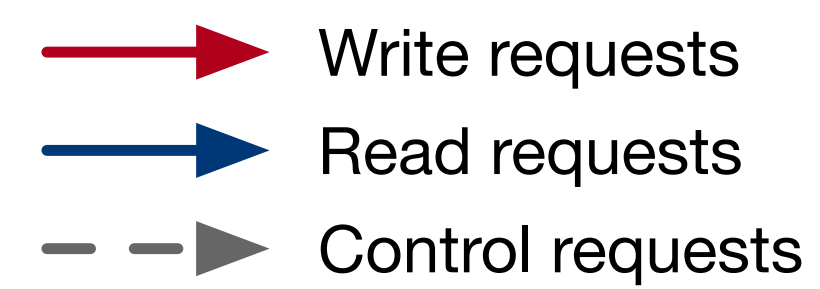




# Problem #6: DynamoDB, again







	<b>DynamoDB</b>	<b>BigTable</b>
<b>99th Percentile Write Latency (ms)</b>	70-100	50-150
<b>99th Percentile Read Latency (ms)</b>	100-2500	~250
<b>LOC</b>	~2000	~400

DynamoDB numbers courtesy of Weaveworks



# Closing thoughts



1. DynamoDB Write Throughput
2. DynamoDB Write Throughput, again
3. Recording rules and alerts
4. Long tail
5. Cost
6. DynamoDB, again



Running for >12months

- Availability: querier unavailable for <12hrs      **~99.9%**
- Durability: lost <2 days of data      **>99.5%**
- 99th percentile write performance      **~60ms**
- 99th percentile query performance      **<200ms**



## Future

- Direct chunk writes from Prometheus to Cortex Chunk Store
- Separate ingester index for better load balancing
- Use prometheus/tsdb for the ingesters
- Etcd & gossip for ring storage
- Chunks in Google Cloud Storage





One more thing...



I left Weaveworks at the beginning of June to focus on Prometheus & Cortex development.

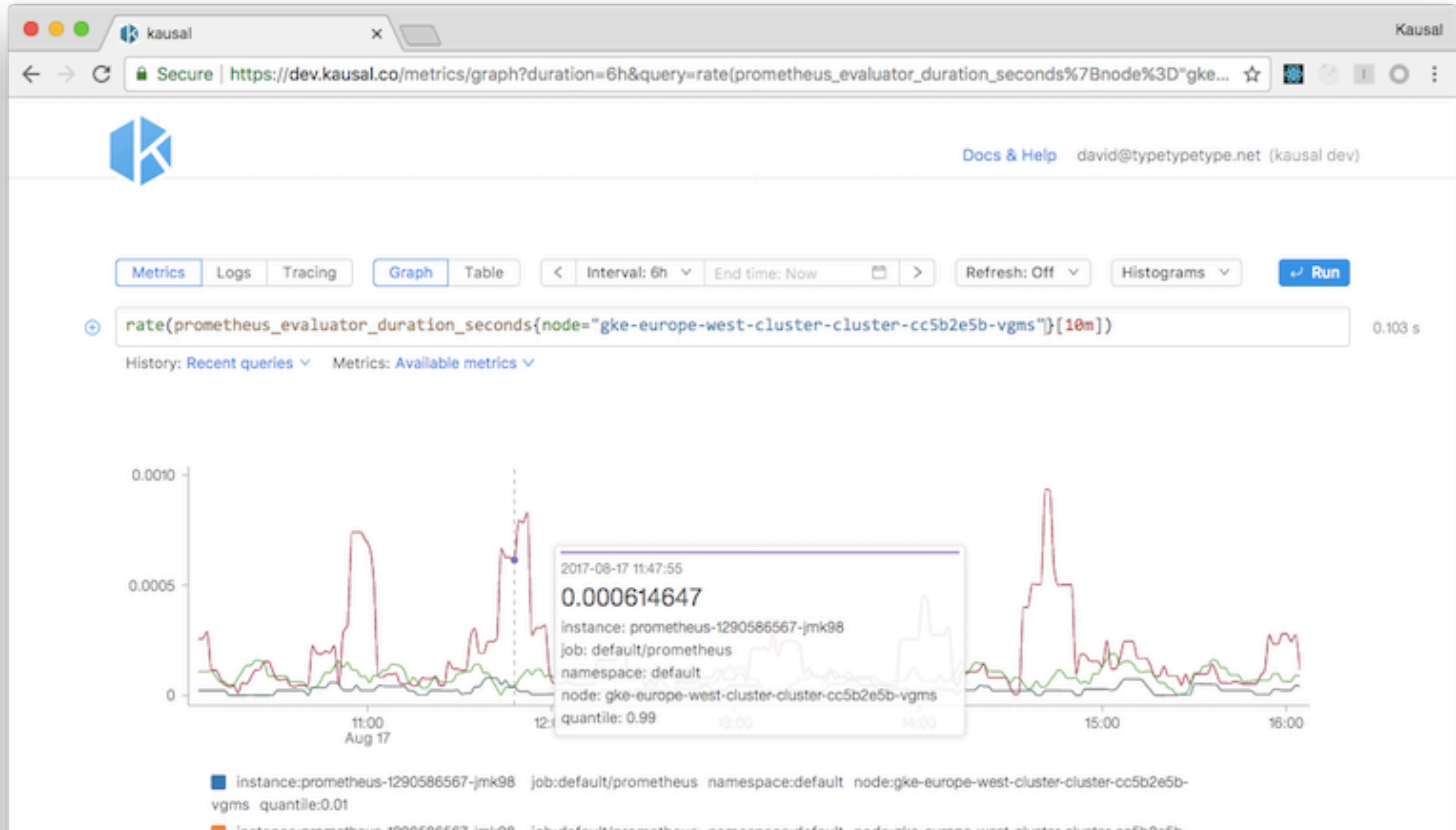
Since then I've teamed up with David to develop some ideas around Prometheus, logging, and tracing.

We're available for Prometheus hosting, consulting, training and support.

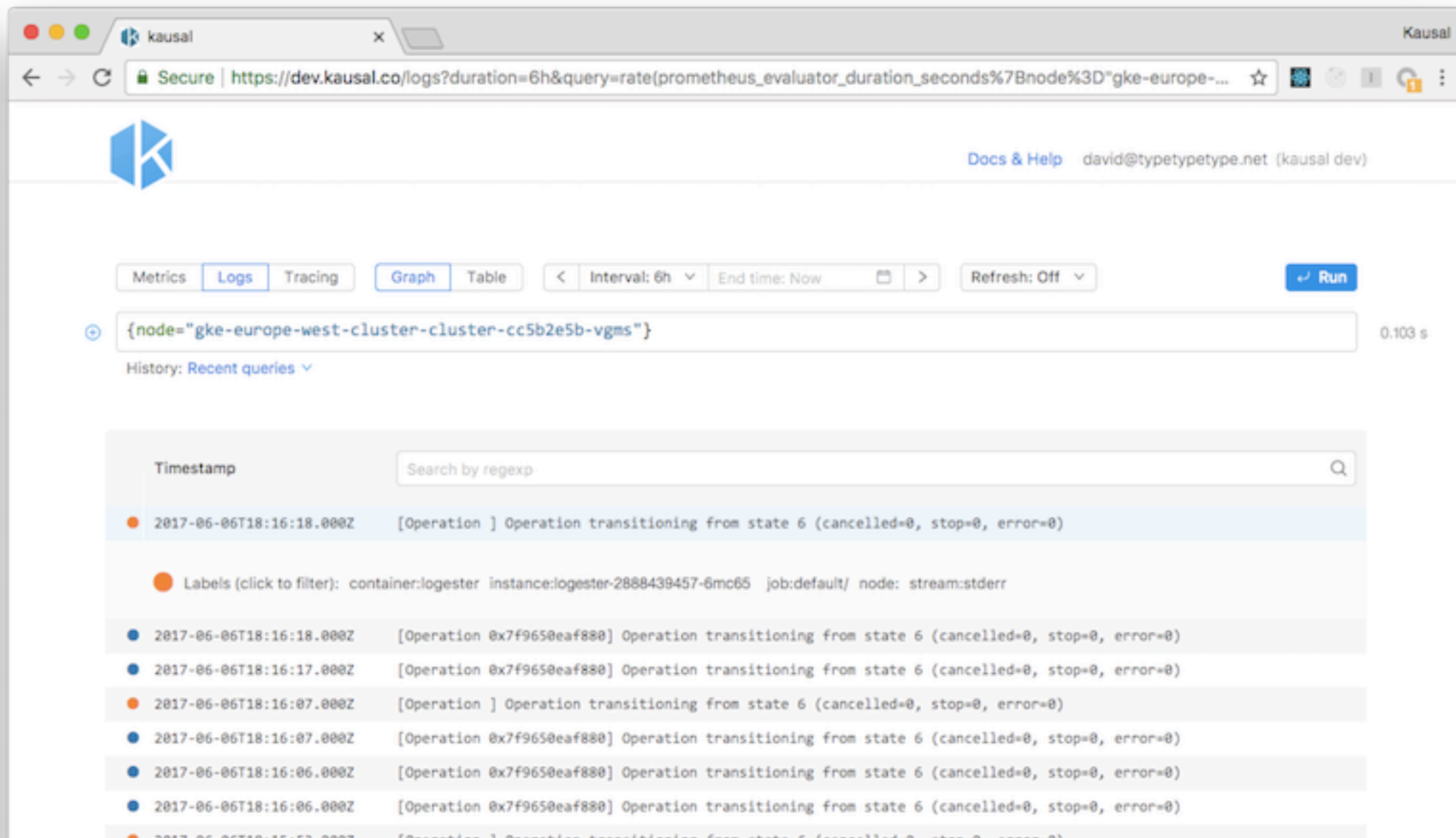
email: [hello@kausal.co](mailto:hello@kausal.co)



# Metrics



# Logs



The screenshot shows the Kausal web interface for viewing logs. The browser address bar shows the URL: `https://dev.kausal.co/logs?duration=6h&query=rate(prometheus_evaluator_duration_seconds%7Bnode%3D"gke-europe-...`. The page header includes the Kausal logo, a "Docs & Help" link, and the user email  `david@typetypetype.net (kausal dev)`.

The main interface features a navigation bar with tabs for "Metrics", "Logs", "Tracing", "Graph", and "Table". The "Logs" tab is selected. Below the navigation bar, there are controls for the query: "Interval: 6h", "End time: Now", and "Refresh: Off". A blue "Run" button is on the right. The query input field contains `{node="gke-europe-west-cluster-cluster-cc5b2e5b-vgms"}` and shows a response time of "0.103 s".

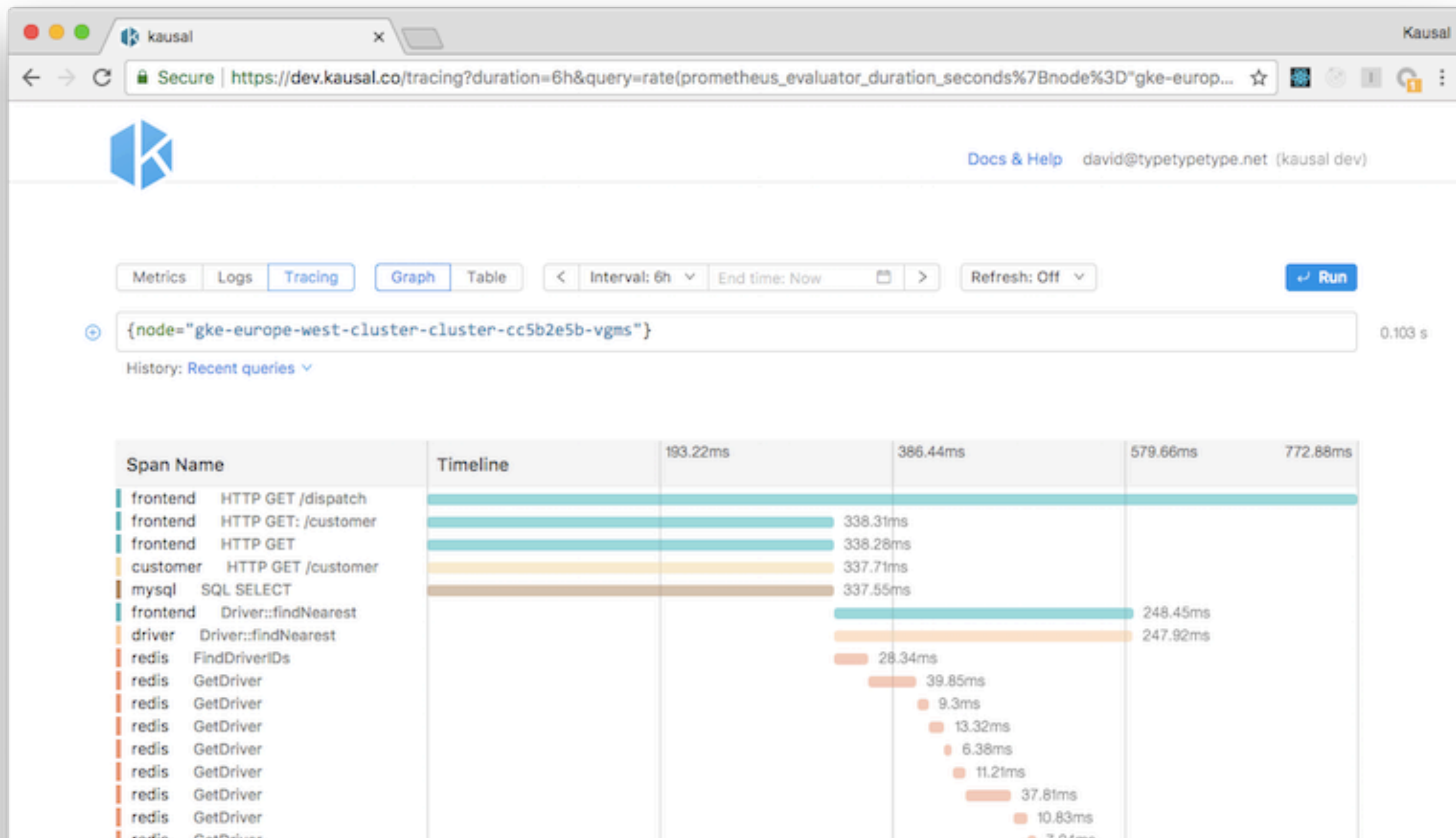
Below the query input, there is a "History: Recent queries" dropdown. The main content area displays a list of log entries. Each entry consists of a colored dot (representing the log level), a timestamp, and the log message. The first entry is highlighted in blue and shows a red dot, indicating an error or warning level. The log message is: `[Operation ] Operation transitioning from state 6 (cancelled=0, stop=0, error=0)`. Below this entry, there are labels: `Labels (click to filter): container:logester instance:logester-2888439457-6mc65 job:default/ node: stream:stderr`. The subsequent entries show blue dots and similar log messages, indicating successful operations.

Timestamp	Log Message
2017-06-06T18:16:18.000Z	[Operation ] Operation transitioning from state 6 (cancelled=0, stop=0, error=0)
2017-06-06T18:16:18.000Z	[Operation 0x7f9650eaf880] Operation transitioning from state 6 (cancelled=0, stop=0, error=0)
2017-06-06T18:16:17.000Z	[Operation 0x7f9650eaf880] Operation transitioning from state 6 (cancelled=0, stop=0, error=0)
2017-06-06T18:16:07.000Z	[Operation ] Operation transitioning from state 6 (cancelled=0, stop=0, error=0)
2017-06-06T18:16:07.000Z	[Operation 0x7f9650eaf880] Operation transitioning from state 6 (cancelled=0, stop=0, error=0)
2017-06-06T18:16:06.000Z	[Operation 0x7f9650eaf880] Operation transitioning from state 6 (cancelled=0, stop=0, error=0)
2017-06-06T18:16:06.000Z	[Operation 0x7f9650eaf880] Operation transitioning from state 6 (cancelled=0, stop=0, error=0)





# Traces



# Thank you!

Questions?

