Multichip Module Packaging and Its Impact on Architecture

by Hubert Harrer

Presentation: CAS/CPMT/CS/SPS Society Chapters

Monday, October 20, 2008   Santa Clara, CA USA

 © 2008 IBM Corporation

IEEE CAS

- This talk is part of IEEE CAS Distinguished Lecturing Series
- Thanks to the IEEE Chapter for inviting and organizing this talk
- Encouraging IEEE membership

© 2008 IBM Corporation

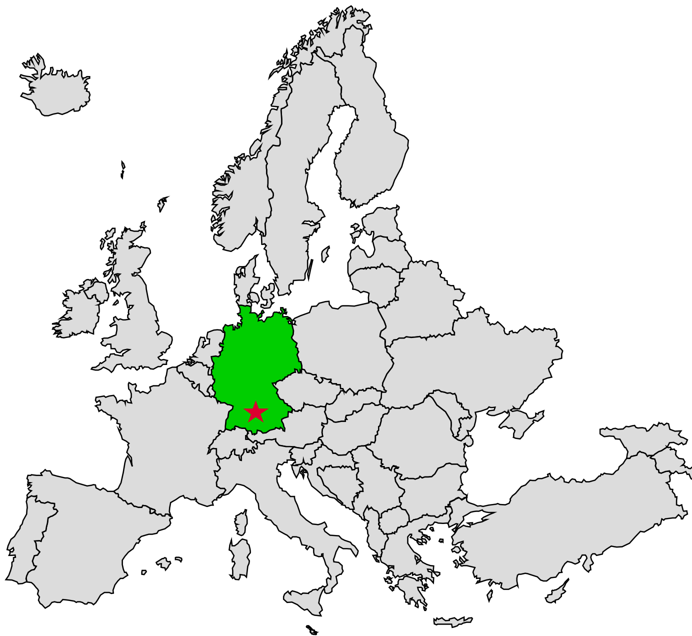
Outline

- Introduction
- System z CEC Architecture Overview
- MCM Technology
- Card Technology
- Bandwidth Requirements
- Power Consumption
- Design Methodologies

© 2008 IBM Corporation

IBM Development Laboratory Boeblingen/Germany

- **About 1900 employees**
  - 500 in hardware
  - 1400 in software



© 2008 IBM Corporation

z-Series System CEC Development

Year	00 z900	02 z900 +	03 z990	05 z9	08 z10
Uni Perf.	0.52	0.64	1.00	1.35	2.16
SMP Perf.	6.19	7.34	20*	40**	60
Processors per MCM	20	20	12-16*** 4 nodes	12-16*** 4 nodes	20**** 4 nodes
Chip Tech.	0.18um	0.18um	0.13um	0.09um	0.065um
Processor Frequency in GHz	0.77	0.91	1.20	1.72	4.40
Package Bitrate in Gb/s	0.39	0.46	0.60	1.72	2.93

*) 8 CPUs per node for workload

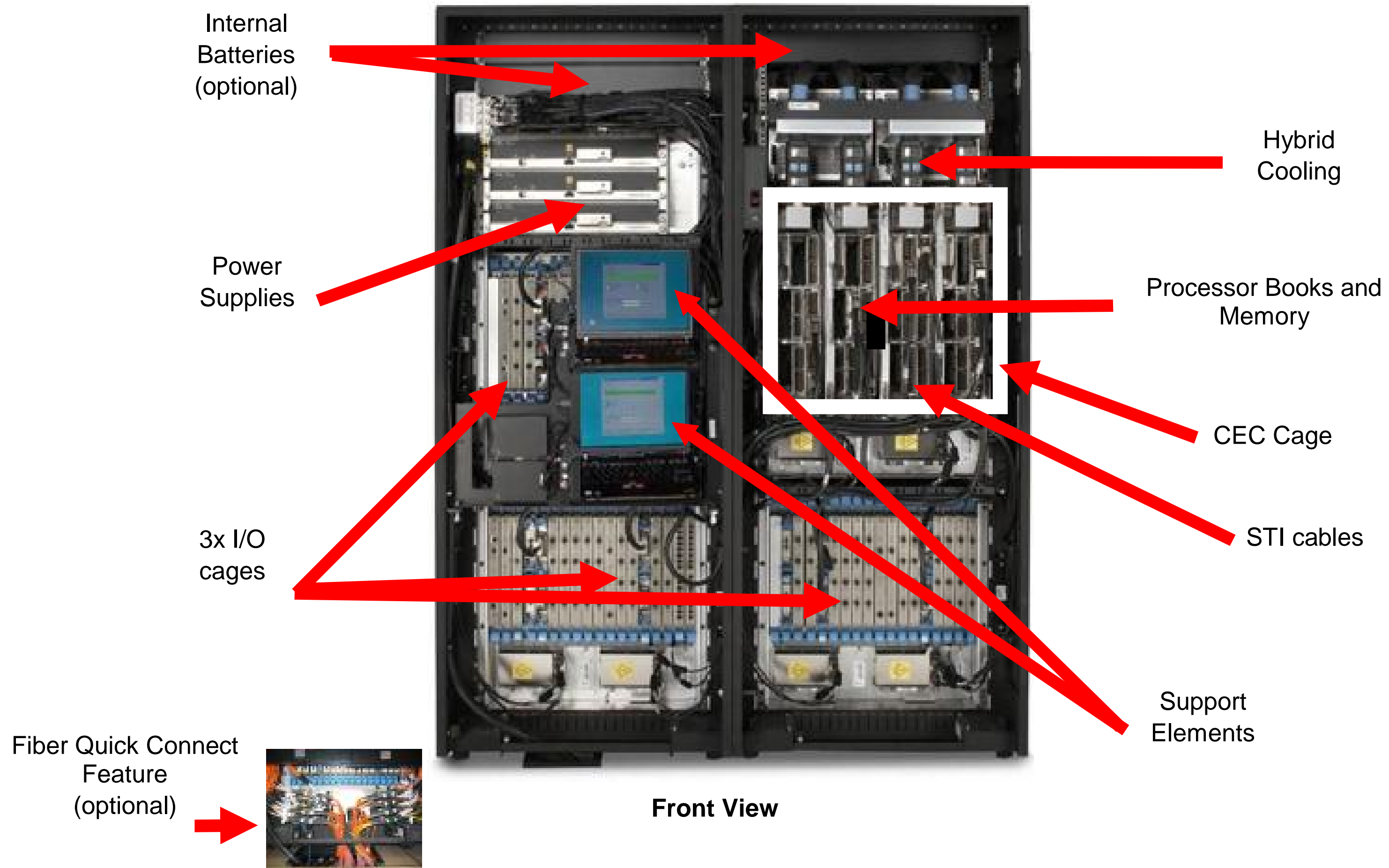
**) 12 CPUs per node for workload

***) dual core processors

****) quad core processors

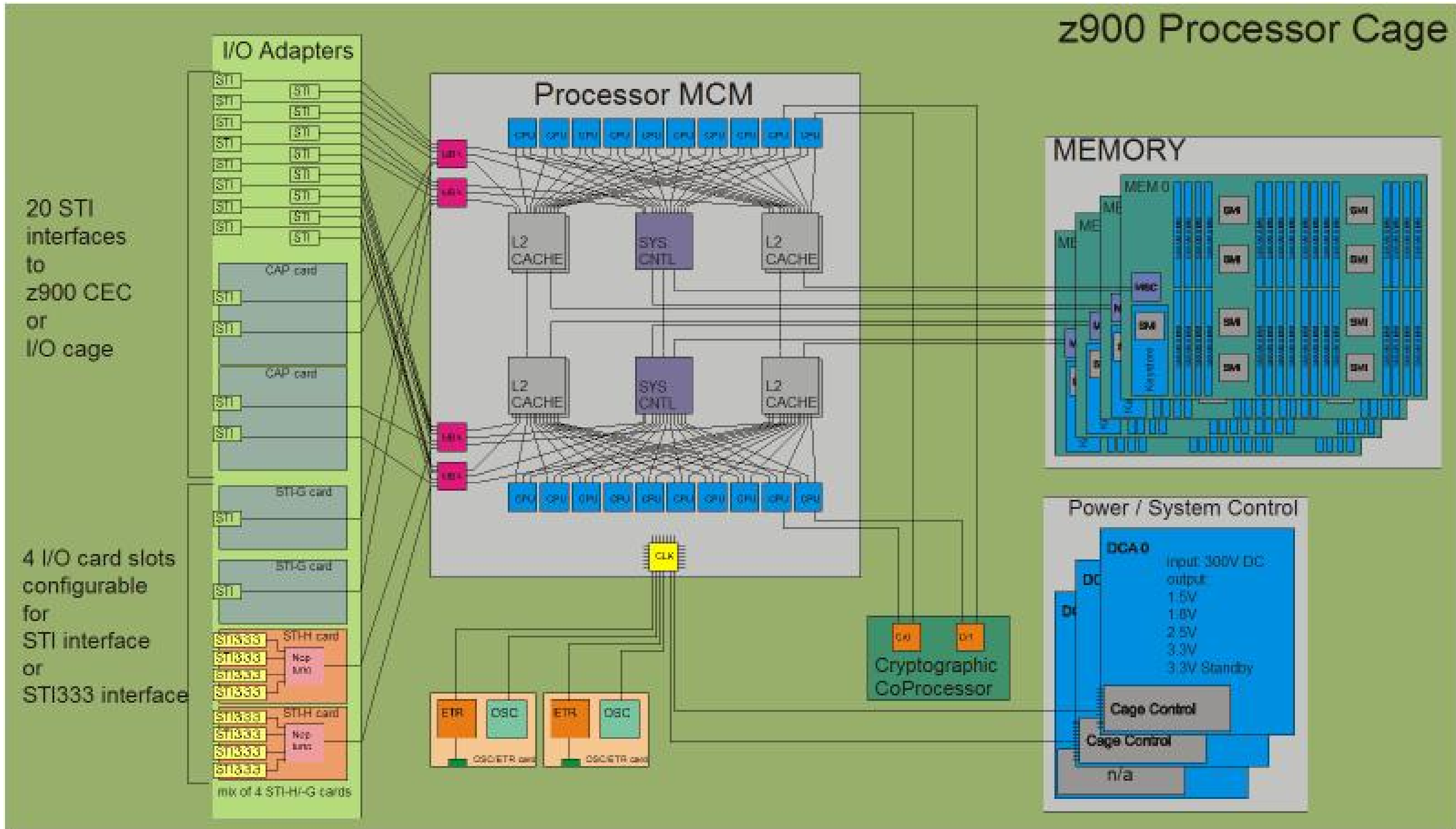
© 2008 IBM Corporation

System z9 Frame



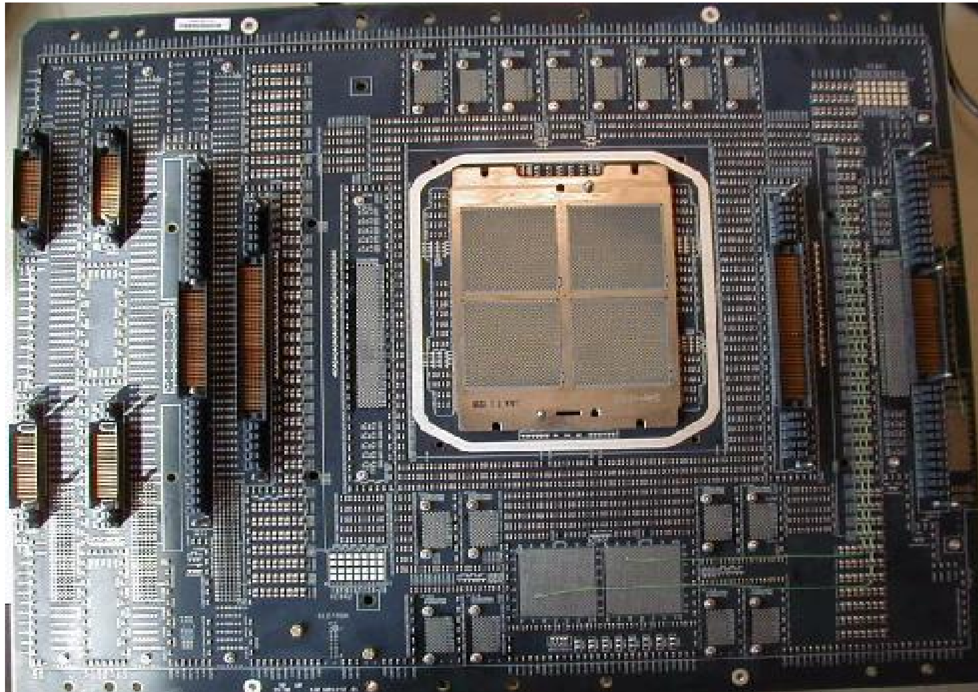
© 2008 IBM Corporation

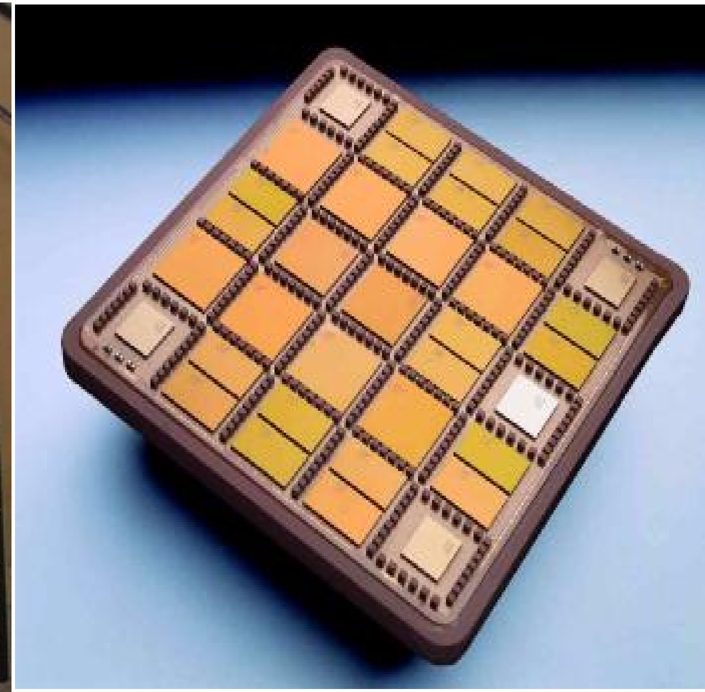
Z900 System Structure



© 2008 IBM Corporation

CEC Processor Board and MCM





- standard loss board material
- LIF (low insertion force) MCM connector 4224 pins (2489 signals)

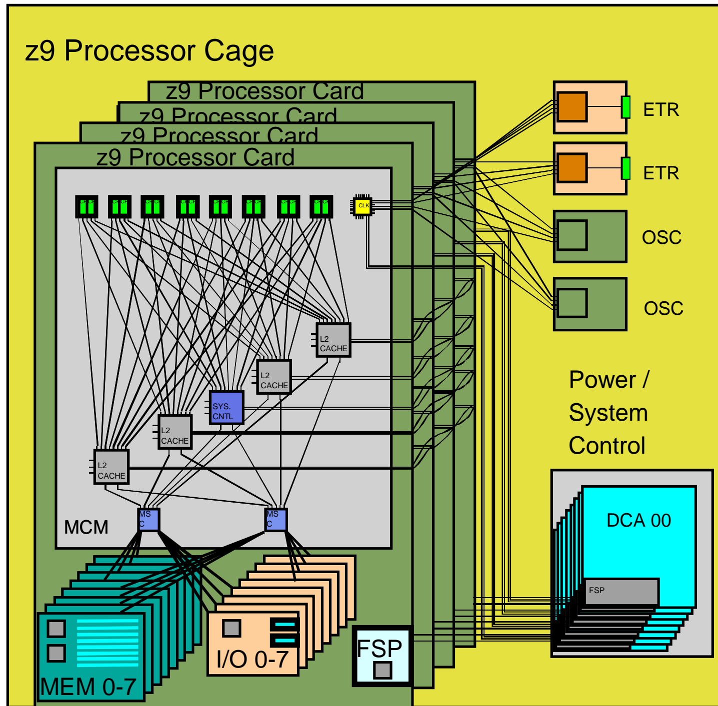
 © 2008 IBM Corporation

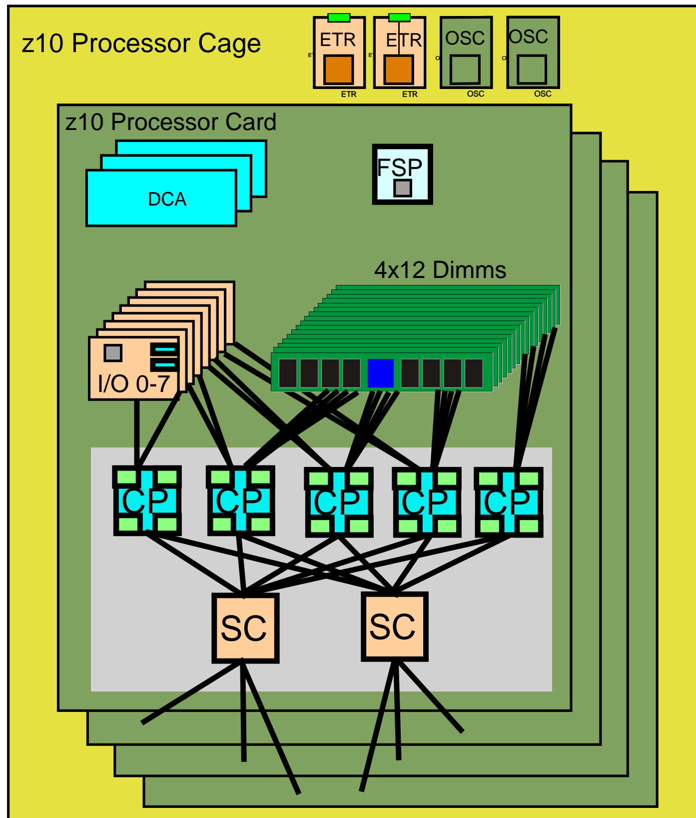
Technical Achievements of Most Complex Multichip Module:

- **Wireability**
  - 1000 m wiring length of 16 k nets
  - 2 x 10 CPs fully connected to 2 x 4 SCDs within two clusters
- **Cooling**
  - 1300 W on 35 chips
  - low temperature cooling of 0 C
- **Cycle Time and Noise Constraints**
  - 459MHz Operation of 80% of the nets (1:2 Gear Ratio) supported by package
- **2 different technologies for implementation**
  - MCM-D technology (IBM)
  - MCM-C technology (Hitachi)

© 2008 IBM Corporation

Logic Structure Comparison between z9 and z10

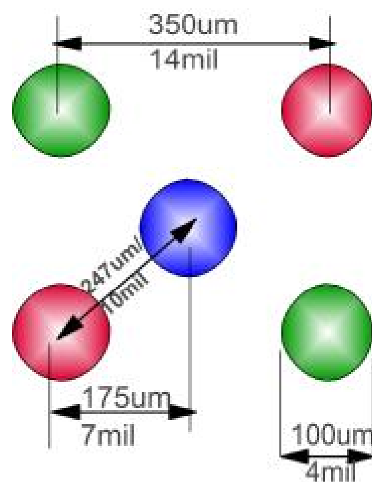




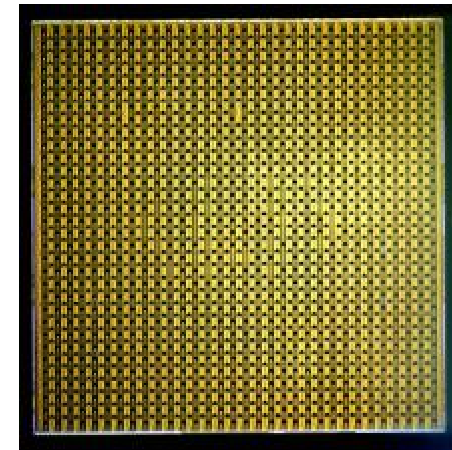
© 2008 IBM Corporation

z9 Chip Technology

Chip	Size [mm*mm]	used signal C4s/Pins
CPU (90nm)	15.78 x 11.84	603
SC (90nm)	16.32 x 16.32	1768
SD (90nm)	15.20 x 15.62	1600
MSC (90nm)	14.24 x 14.24	1344
CLK (120nm)	9.40 x 9.48	487
MCM	95 x 95	2923



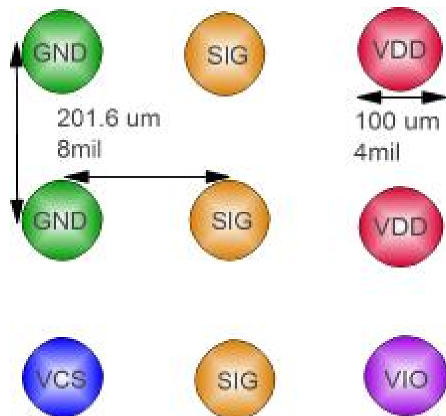
C4 Structure:
Controlled Collapsed Chip Connect
PbSn balls



© 2008 IBM Corporation

z10 Chip Technology (65nm)

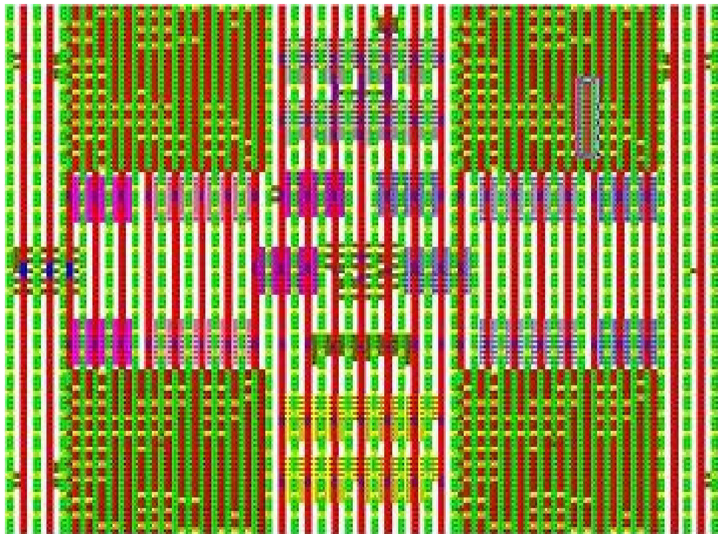
Chip	Size [mm*mm]	used signal C4s/Pins
CPU (65nm)	21.8 x 21.1	1180
SC (65nm)	21.6 x 21.0	2411
MCM	96 x 96	3490

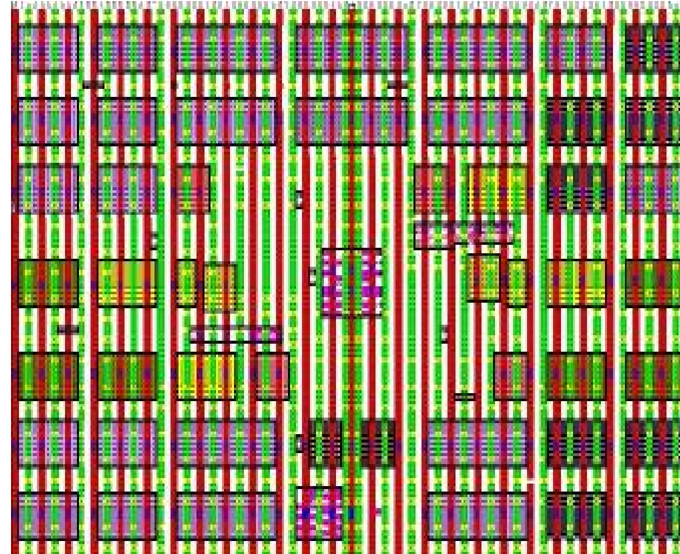


C4 Structure

© 2008 IBM Corporation

Chip Footprint





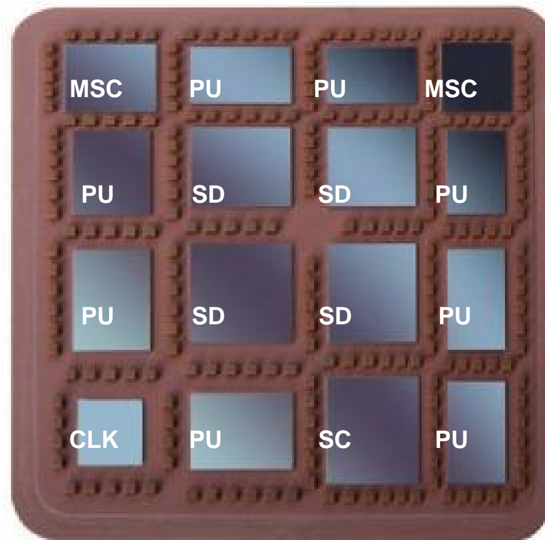
CP	
GND	3394
VDD	3171
VCS	642
VIO	191
Signal	1180
TOT	8587

SC	
GND	2166
VDD	2483
VCS	654
VIO	346
Signal	2411
TOT	7976

© 2008 IBM Corporation

z9-Multi Chip Module (MCM)

- **Advanced 95mm x 95mm MCM**
  - 102 Glass Ceramic layers
  - 16 chip sites, 217 capacitors
  - 0.545 km of internal wire

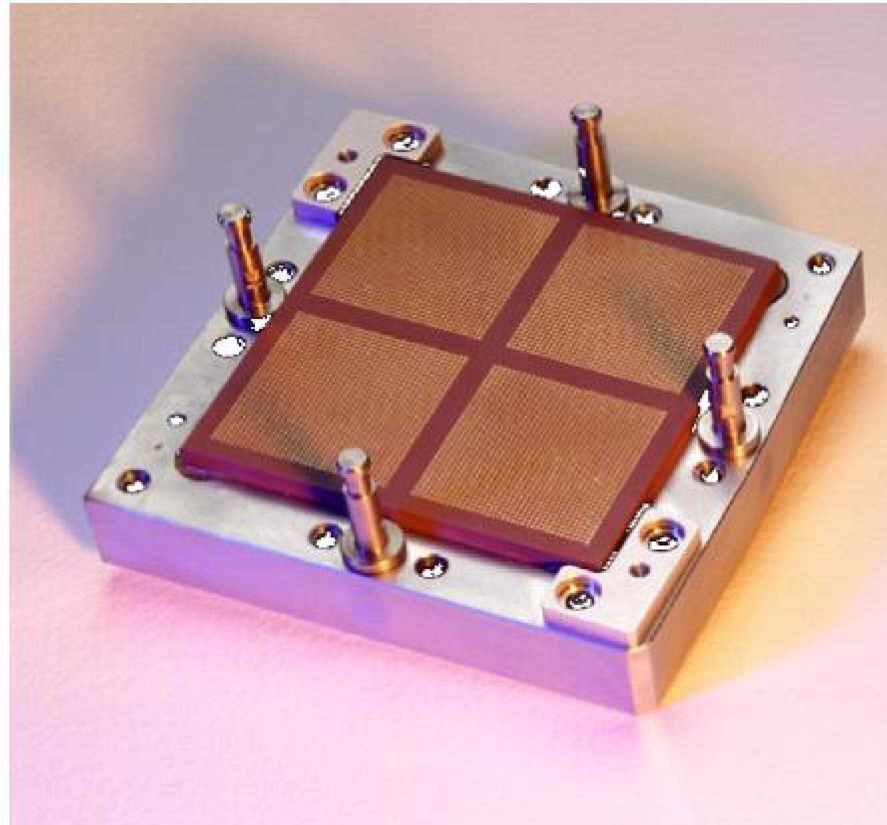


- **CMOS 90nm chip Technology**
  - PU, SC, SD and MSC chips
  - Copper interconnections, 10 copper layers
  - 8 PU chips/MCM
    - 15.78 mm x 11.84 mm
    - 121 million transistors/chip
    - L1 cache/PU
      - 256 KB I-cache
      - 256 KB D-cache
    - 0.58 ns Cycle Time
  - 4 System Data (SD) cache chips/MCM
    - 15.66 mm x 15.40mm
    - L2 cache per Book
      - 660 million transistors/chip
      - 40 MB
  - One Storage Control (SC) chip
    - 16.41mm x 16.41mm
    - 162 million transistors
    - L2 cache crosspoint switch
    - L2 access rings to/from other MCMs
  - Two Memory Storage Control (MSC) chips
    - 14.31 mm x 14.31 mm
    - 24 million transistors/chip
    - Memory cards (L3) interface to L2
    - L2 access to/from MBAs (off MCM)
  - One Clock (CLK) chip - CMOS 8S
    - Clock and ETR Receiver

© 2008 IBM Corporation

LGA Connector

- 5184 LGA (Land Grid Array Connector)
- 4 quadrants with 36 by 36 pads
- 30-50mg per contact
- about 200kg per connector

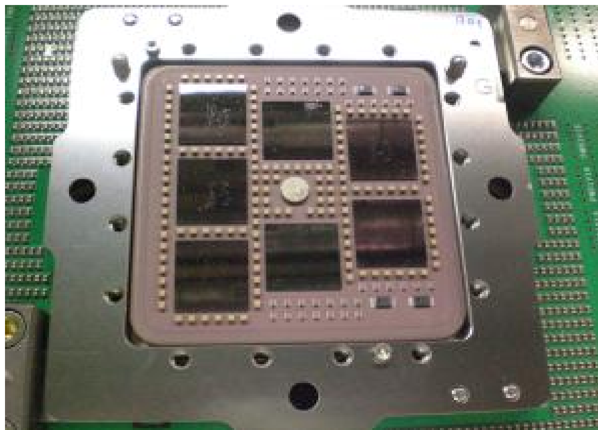


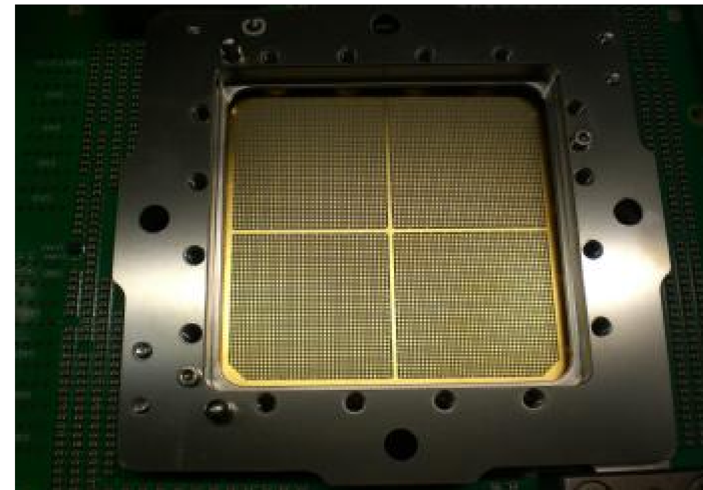
© 2008 IBM Corporation

z10-MCM

§**Advanced 96mm x 96mm MCM**

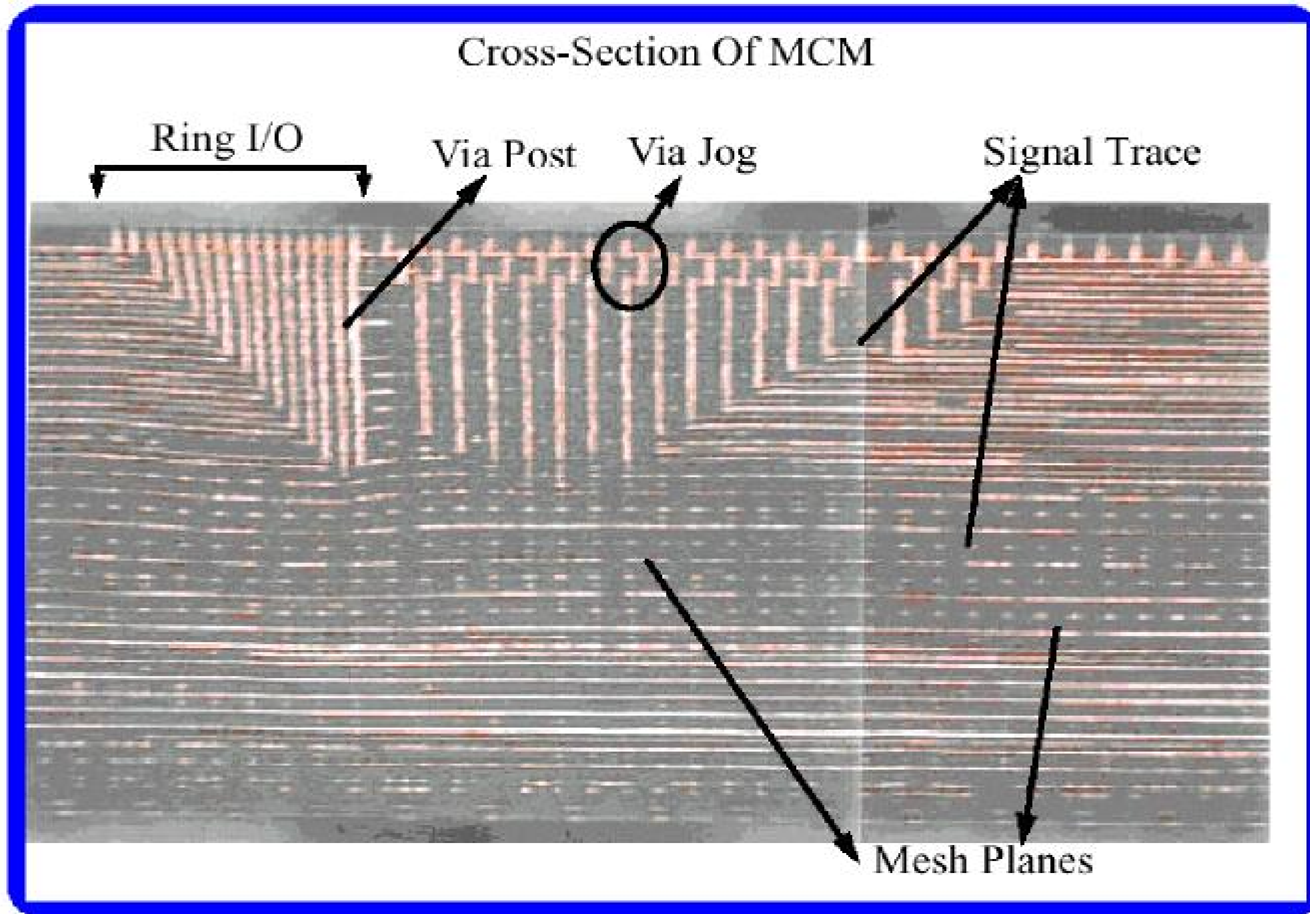
- ►105 Glass Ceramic layers
- ►201.6um pitch
- ►7 chip sites
- ►178 capacitors (138 600nF LICA, 40 1uF IDC)
- ►4 SEEPROMS
- ►6740 nets
- ►0.7 km of internal wire
- ►7356 LGA 1mm pitch
- ►13 different voltage domains
  - –separate VDD and array voltage per CP chip
  - –shared VDD and array voltage for both SC chips
- ►3 additional standby voltages





© 2008 IBM Corporation

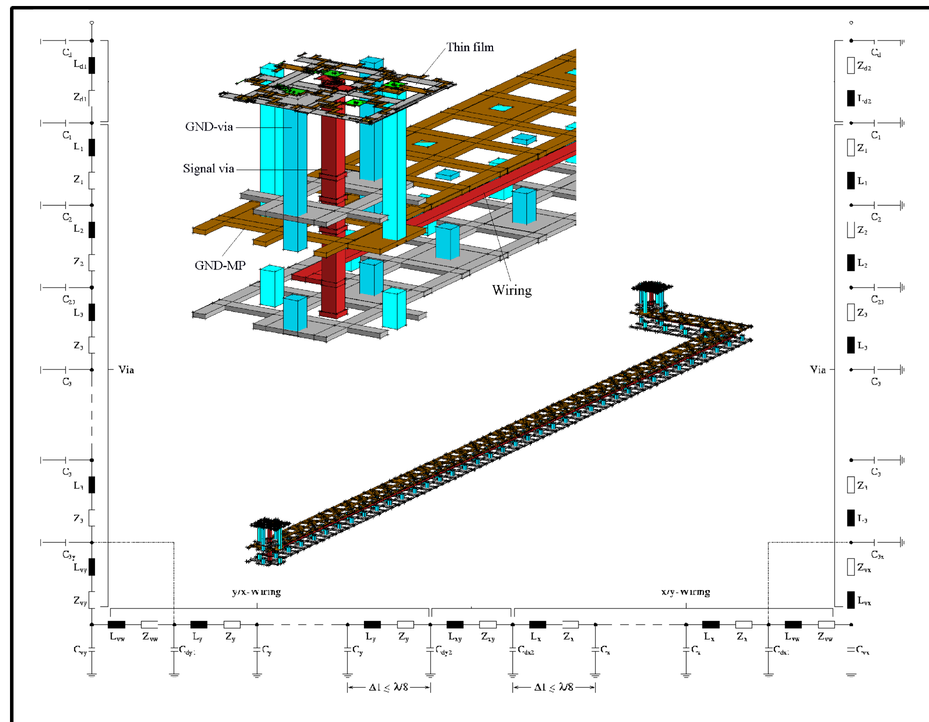
MCM Cross Section



© 2008 IBM Corporation

3D Modeling

- frequency dependent 3d model extraction
- correction factor for actual manufacturing shapes with 2d tools
- in future direct s-parameter extraction from 3d geometries

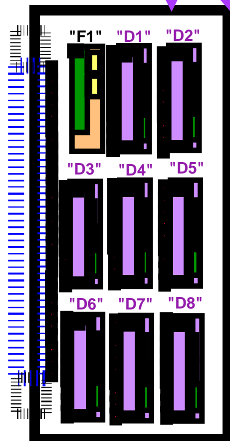


 © 2008 IBM Corporation

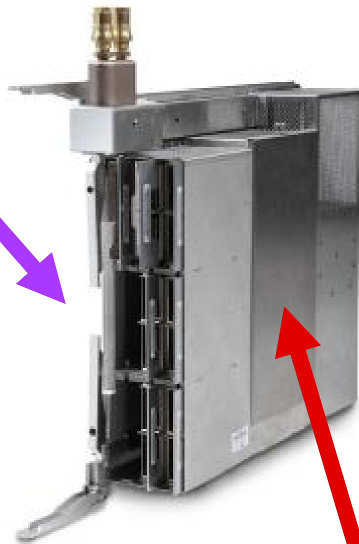
Processor Book Layout

Processor Node Card
- 378mm x 470mm
- 3740 nets, 645m
- cross section 10S21P2MP
- standard loss material, buried vias
- 180 1.5mF decaps for low frequency decoupling
- 1604 10uF ceramic decaps
- 18733 PTH
- 1160 VHDM connector for ring

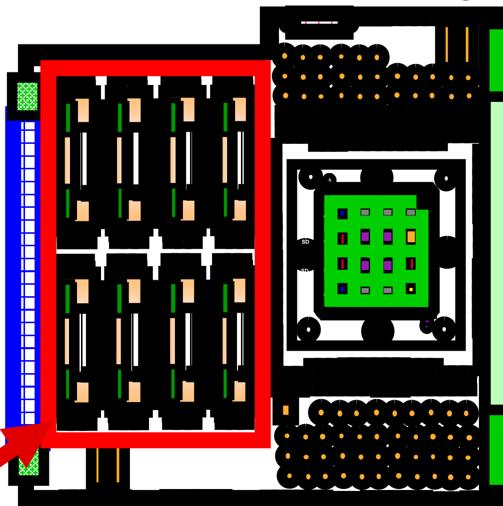
Up to 8
hot pluggable MBA/STI
fanout cards



Front View



Memory Cards
Up to 128 GB

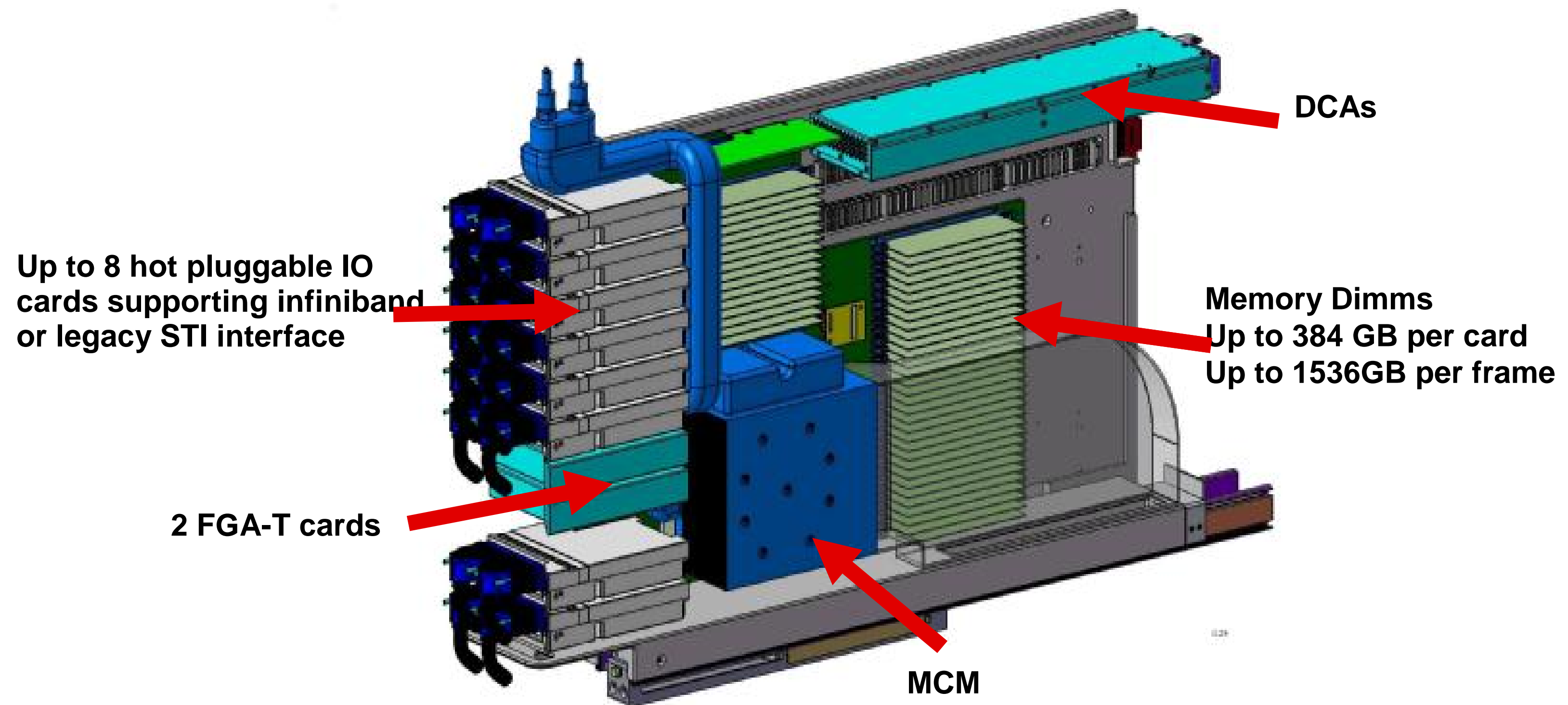


Side View

© 2008 IBM Corporation

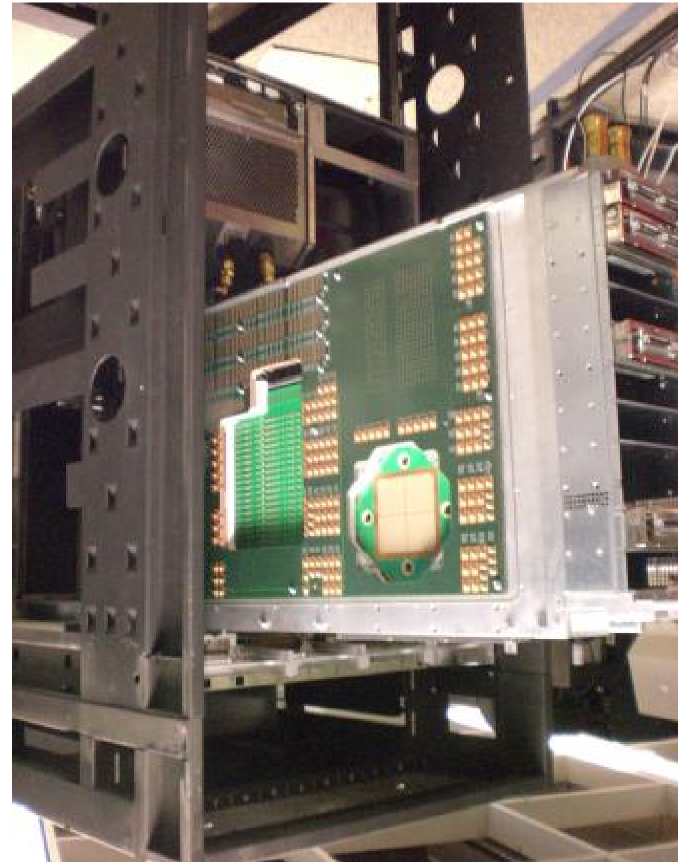
Processor Book Layout
or
The Mother of a Blade

Processor Node Card
- 584mm x 460mm
- 5330 nets, 864m
- cross section 10S18P2MP
- low loss material, buried vias
- 2523 decaps
- 29906 PTH, 4614 buried vias
- 14 row Ventura connector for fabric with 1680 signals



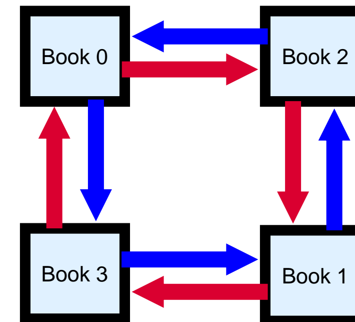
© 2008 IBM Corporation

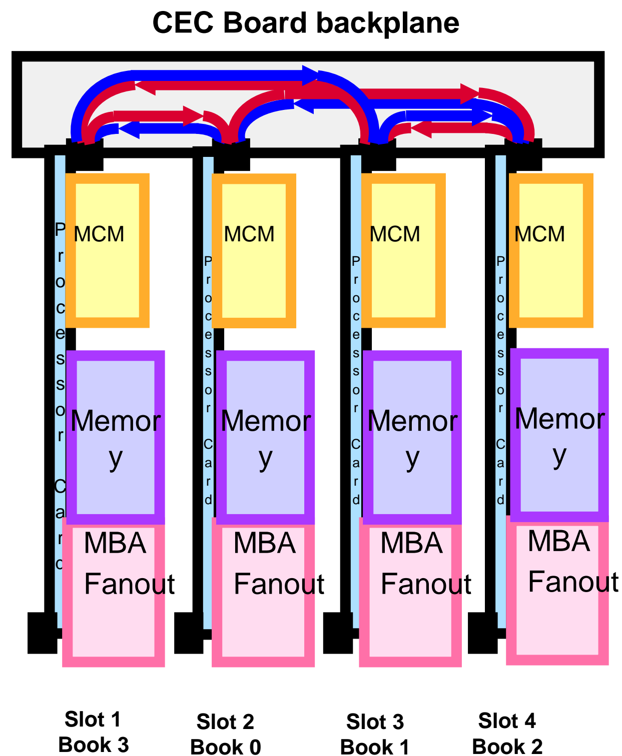
Actual Node Picture



 © 2008 IBM Corporation

Communication Ring Structure



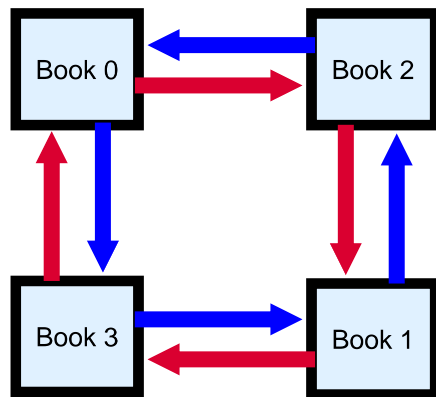


- **The ring structure consists of two rings (one running clockwise, the other counterclockwise)**
- **In a two or three Book configuration, Jumper Book(s) (installed in the CEC cage) complete the ring**
  - Jumper Books are not needed for a single-Book configuration
- **Books may be able to be inserted into or removed from the ring nondisruptively**
  - May allow **Concurrent book add** for model upgrade
  - **Enhanced book availability** to return a book after removal for upgrade or repair

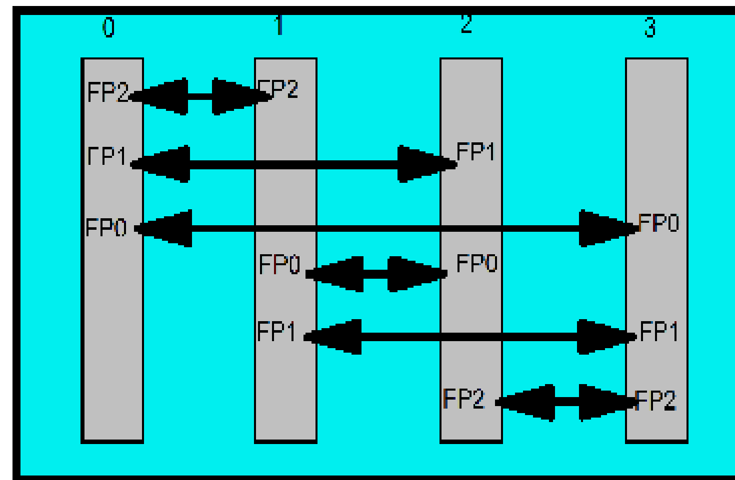
© 2008 IBM Corporation

Communication between Nodes

§lower latency with direct fabric structure instead of ring



z9



z10

© 2008 IBM Corporation

Cooling and Power

§**Cooling**

- ►Improved Small Gap Technology (3.5+-1mil)
- ►Tim1: 26/39 C/W/mm^2
- ►Tim2: 20 C/W/mm^2

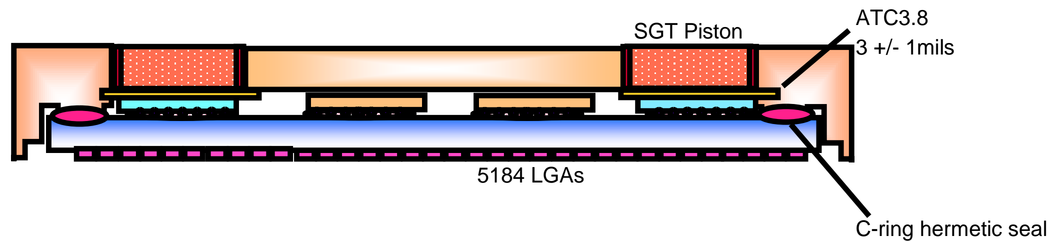
§**Power**

- ►Frame Power: 17.2kW
- ►Node Power: 3.5kW

§**Temperature**

- ►45C chip junction

z9 MRU cooling
with aircooled backup mode



© 2008 IBM Corporation

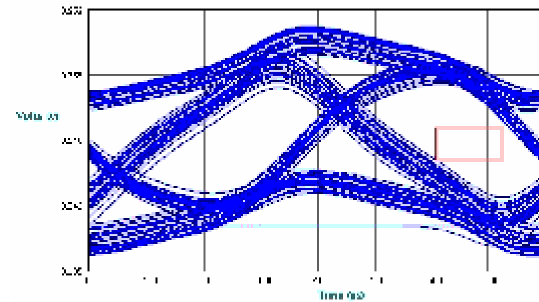
Test Concept

- **Wafer Test**
  - selftest, speed characterization
- **Temporary Chip Attachment on Single Chip Module**
- **Burn In (high voltage, high temperature)**
  - 120C, 1.4 x voltage for 24 hours
- **MCM Assembly Test**
- **System test in a 2 Node Configuration**
  - slower /faster cycle time
  - lower/higher voltage
  - nominal /higher temperature
- **System test in a 4 Node Configuration**
- **Final System Test in Customer Configuration**

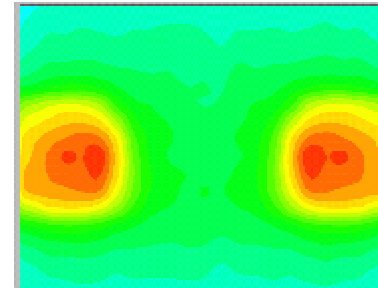
© 2008 IBM Corporation

Packaging Key Challenges

- **Bus Bandwidths**
  - Large increase of bandwidths driven by multicore processors
    - increasing frequency
    - increasing bit lines



- **Power**
  - high end servers will use high power processor chips with increased chip sizes
  - power consumption is limiting performance
    - power saving concepts
    - new cooling concepts



© 2008 IBM Corporation

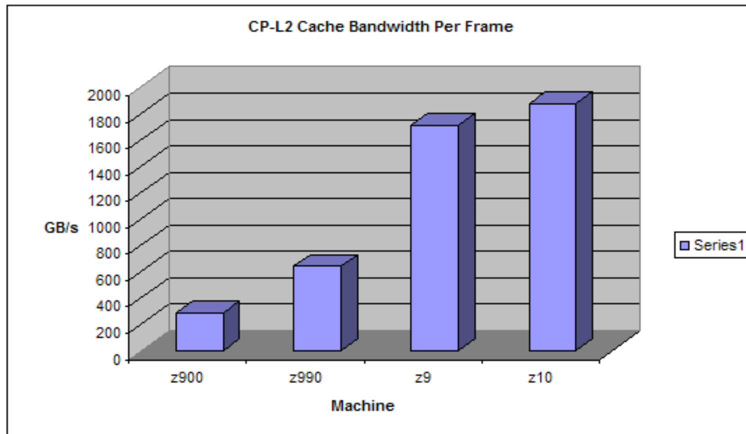
z9/z10 Timing and Bus Performance Comparison

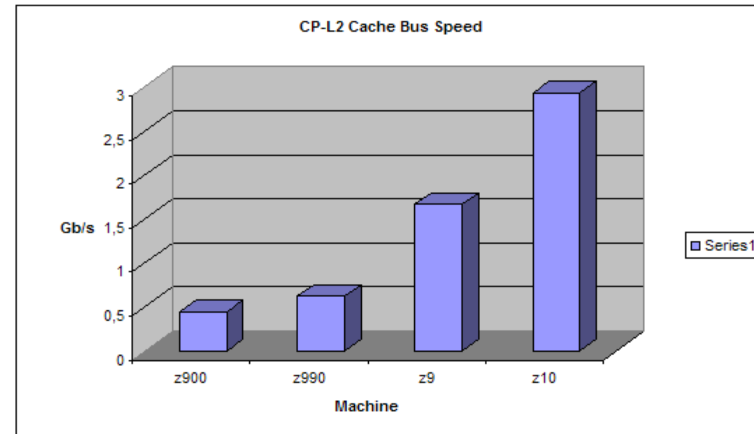
Name	Data Size B	Bit Rate Gb/s	Bandwidth GB/s 4 nodes	1st level Wire	2nd Level Wire	Comment
Cache Bus	256/160 x4	1.72/2.93	1764/1877	13 cm	n.a.	DDR source synch
Memory	64/48 x4	0.86/2.17	220/417	4 cm	20 in	DDR source synch
I/O Hub (MBA)	64/64 x4	0.86/2.2	220/563	3 cm	20 in	DDR source synch
Ring/Fabric (SMP)	32/24 x4	0.86/1.47	110/140	cm	80cm	DDR source synch

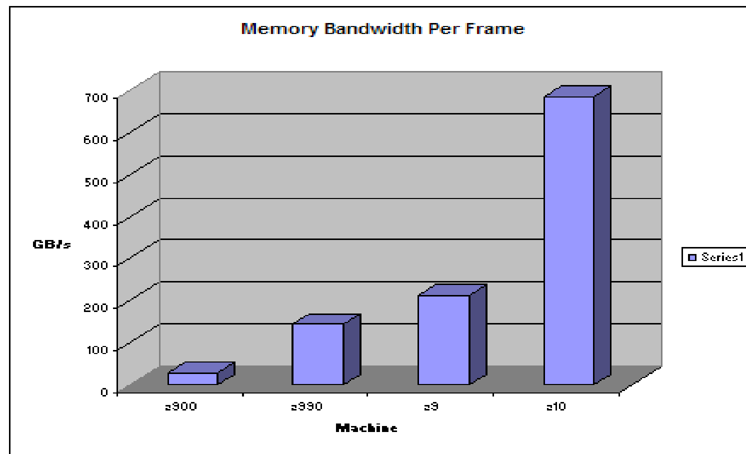
- wiring rule for each net over all components
- system timing run for each net (~12k) with in house tool
- single ended source terminated driver and diode clamped receiver for all nets besides STI

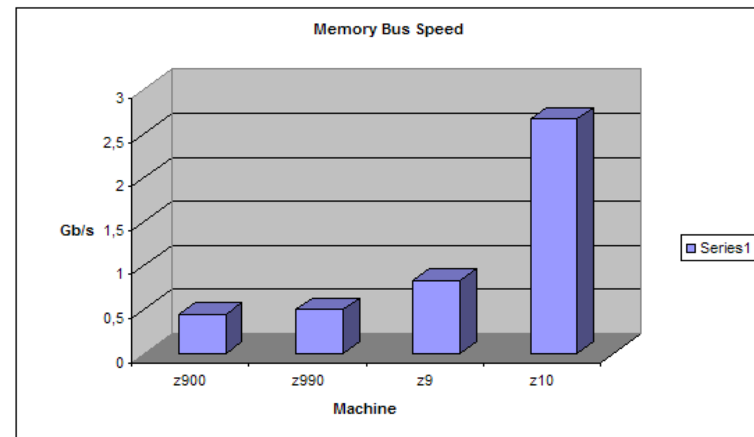
© 2008 IBM Corporation

Bandwith Comparison



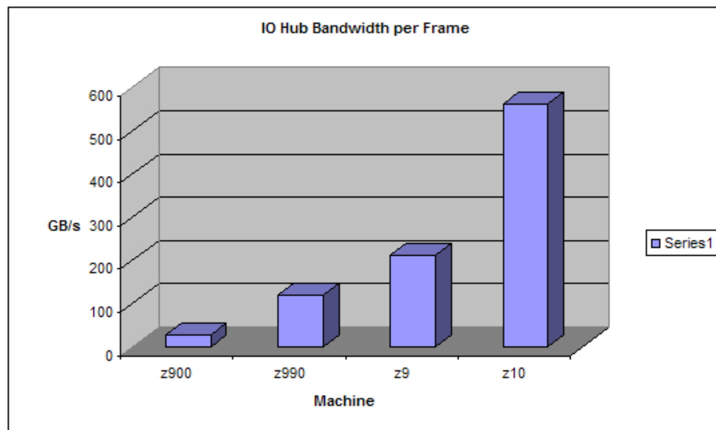


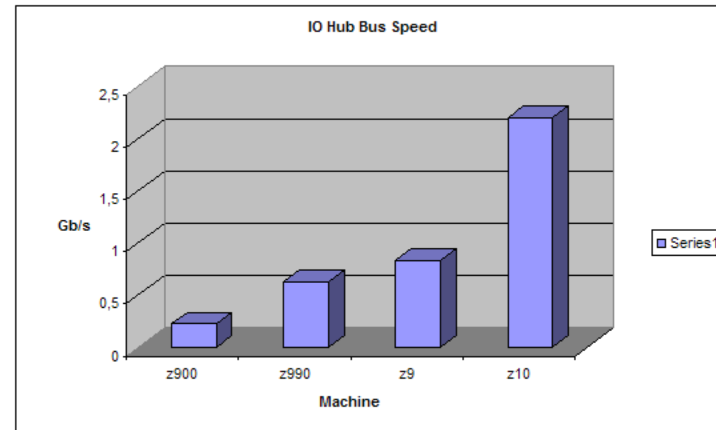


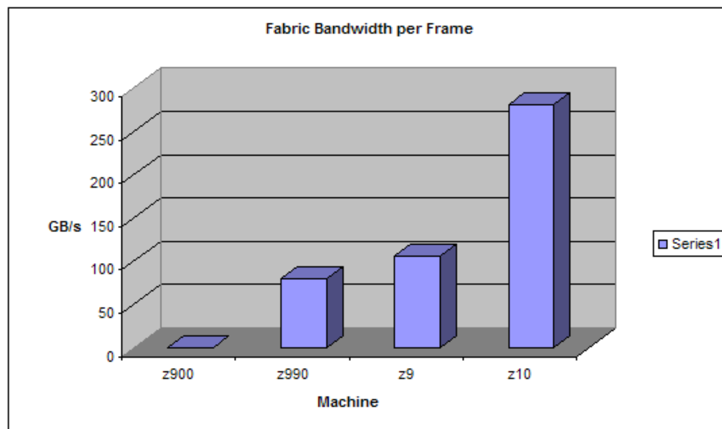


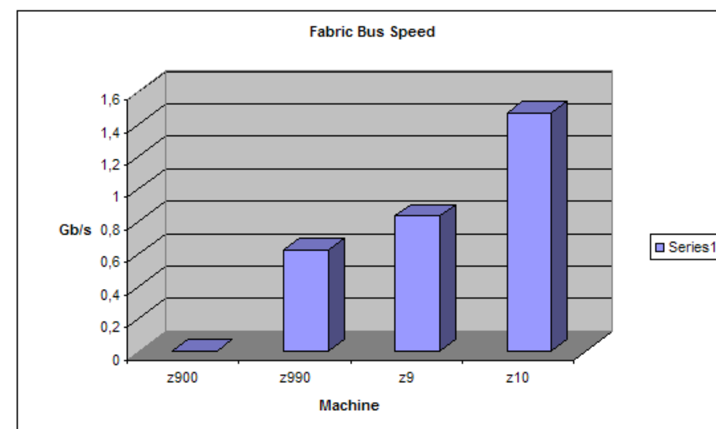
© 2008 IBM Corporation

Bandwith Comparison









© 2008 IBM Corporation

Elastic Interface 3

- **Source Synchronuous Interface**
  - Clocks are transferred together with data
- **Deskewing of wiring lengths during power on sequency**
  - state machine initializes delay elements
- **Elasticity due to buffering receiving data in multiple latch stages**
  - data can be stored on the wire
- **Diagnostic capabilities for bringup and manufacturing test**
  - automated measurement of eye sizes at receiver output
- **Automatic repair capability by including redundant lines**
  - logical rerouting during power on

© 2008 IBM Corporation

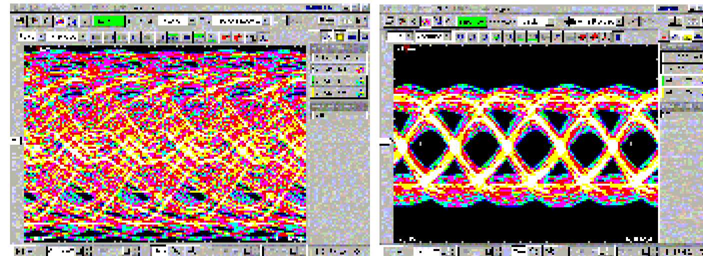
Maximum Data Rates

§Net Topology

	Power	Area	Netlength	Speed	Cost
Single Ended	low	low	short	low	low
Differential	high	large	long	high	low
Optical	high	large	very long	very high	high

§Driver/Receiver Circuits

- ►pre compensation (Driver)
- ►signal restoring (Receiver)

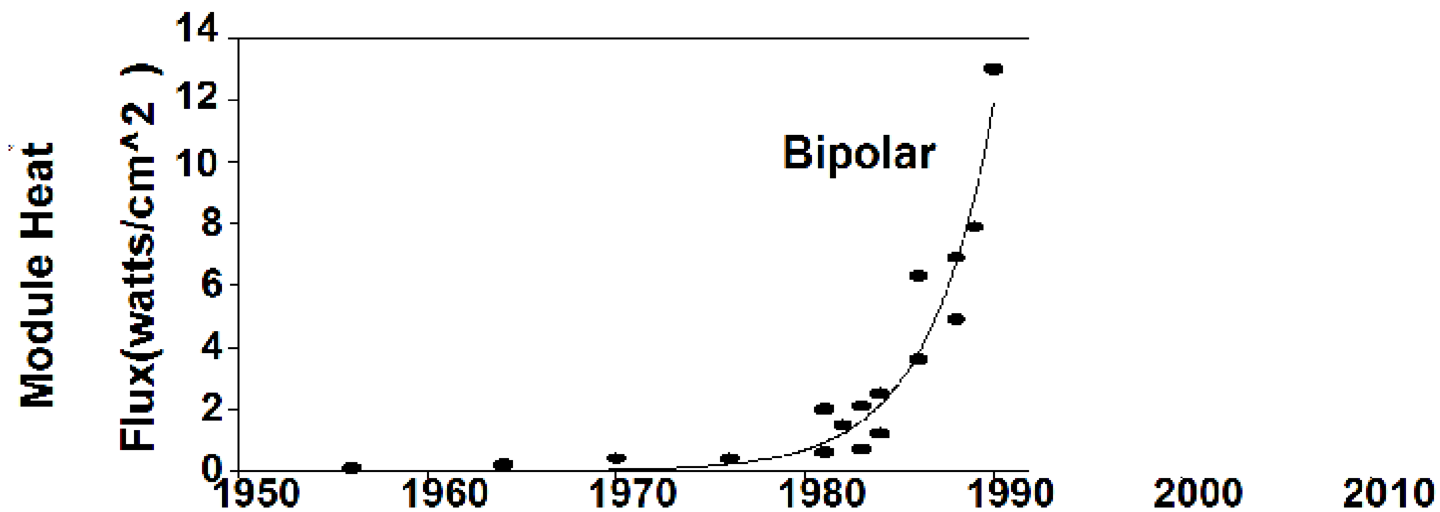


§Channel

- ►Attenuation (dc resistance, skin effect, dielectric loss)
- ►Reflection (characteristic impedance and distortions from vias, connectors)
- ►Crosstalk (line coupling, via coupling, connector coupling)

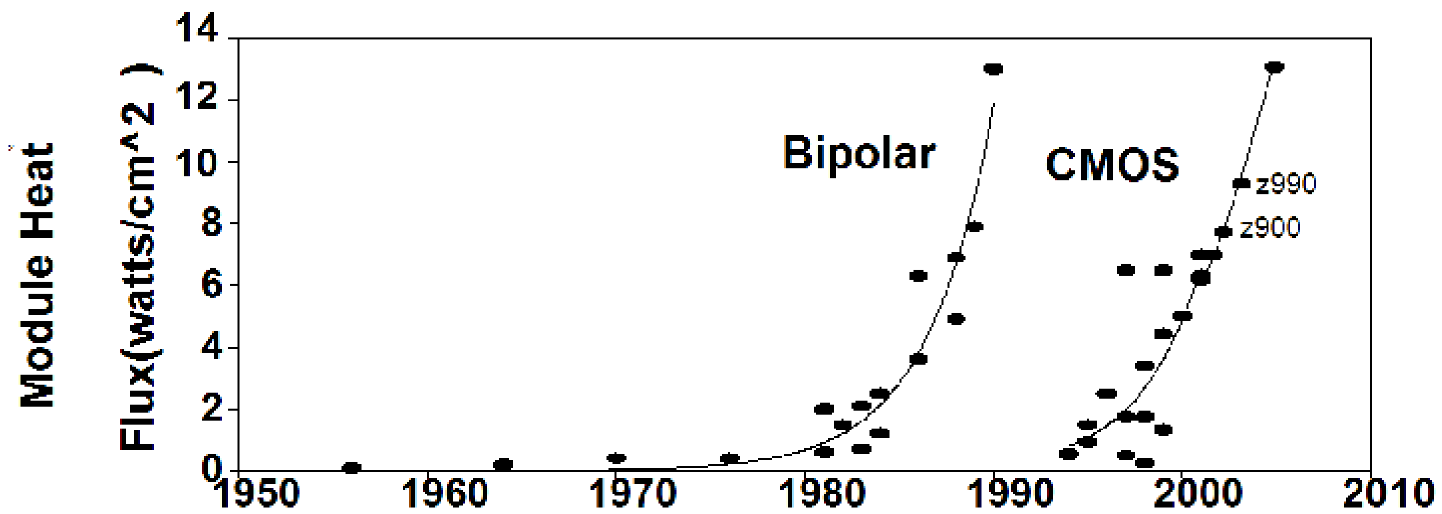
© 2008 IBM Corporation

Comparison of Bipolar and CMOS Power



© 2008 IBM Corporation

Comparison of Bipolar and CMOS Power

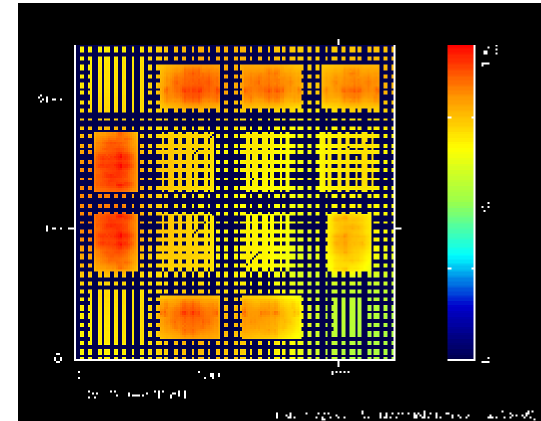


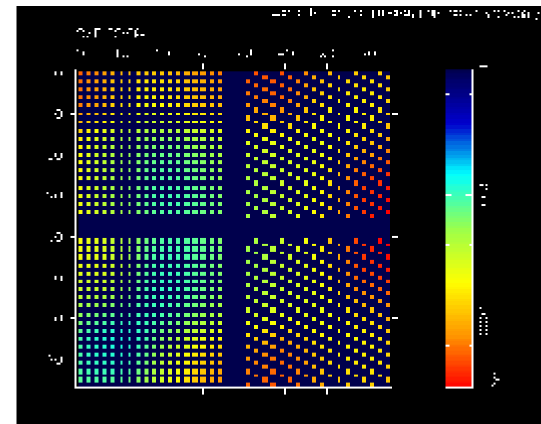
© 2008 IBM Corporation

High Power Challenges

§**Power Delivery**

- ►maximum DC drop (MCM drop 100mV)
- ►maximum Connector Current (LGA 2.0A)
- ►maximum C4 current (200mA)
- ►Current Delivery on Card and Board





© 2008 IBM Corporation

MCM Power

	VDD=1.05V	Logic Power 521W	VDD=1.1V	Logic Power 572W
	T=49C	T=90C	T=49C	T=90C
-5 sigma	1192 W	1926 W	1406 W	2313 W
-4 sigma	980 W	1448 W	1142 W	1722 W
-3 sigma	839 W	1128 W	967 W	1326 W
-2 sigma	759 W	947 W	869 W	1101 W
-1 sigma	721 W	860 W	822 W	993 W

- tradeoff between worst case power and process tolerances
- air cooled backup mode results in highest power due to leakage

© 2008 IBM Corporation

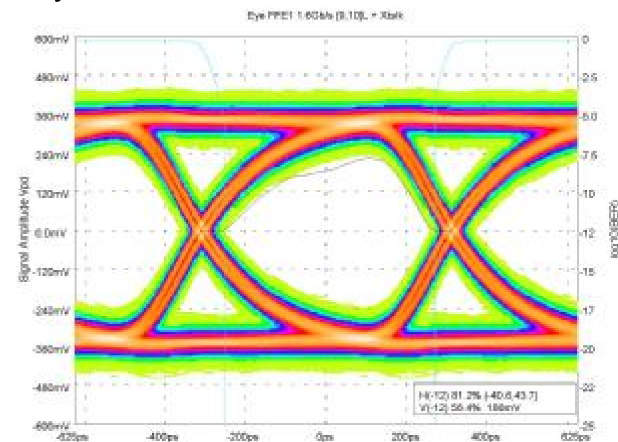
Packaging Design Methodologies

- Pin Optimization over multiple packaging components (IBM tools)
- Placement and Routing Studies (Cadence)
- Constraint Manager (Cadence) to implement rule requirements
- PD in Allegro

© 2008 IBM Corporation

Packaging Design Methodologies

- **Timing and Noise Methodology**
  - **Pre PD:**
    - create 3d models for all components (lines, vias, connectors)
    - power spice simulation
      - use length estimates from wiring analysis
      - use coupling estimates from previous system experience
    - HSSCDR (eye timing tool calculating bit errors probability) (IBM tool)
      - uses s-parameters for channel working in frequency domain
      - linear model of driver
      - driver jitter assumptions included
  - **Post PD:**
    - Fastline simulation for timing and noise (IBM Tool)
    - all nets in the system are being simulated



© 2008 IBM Corporation

Packaging Design Methodologies

- **Power Delivery**
  - First and second level packaging (Pre PD and Post PD)
    - resistive grid analysis with IBM tool
      - based on high level syntax language
      - based on Allegro input
    - system level power simulation
      - voltage drops
      - number of power planes
      - thickness of power planes
      - placement of power connections

© 2008 IBM Corporation

Packaging Design Methodologies

§**Midfrequency Noise**

►Decoupling capacitor simulations with vendor tools

–input on power, switching activity and on chip power map

–PrePD and Post PD simulation

–model to hardware correlation

–multi core power impact

►Hierarchical concept on voltage stabilization

–on chip decaps for HF noise close to hot spots

–MF decaps on 1st level package to load on chip decaps

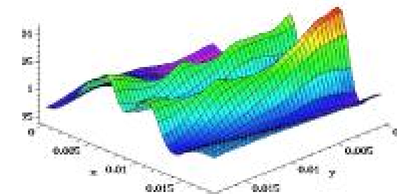
–LF decaps on 2nd level packaging to suppoprt regulation loop

–loading time constant must be designed to guarantee the current loading of the next stage

© 2008 IBM Corporation

Signal Integrity and Decoupling Strategy

	Noise Amplitude	Placement of Capacitors	Type of Capacitors	Number of Capacitors	Amount of Decoupling 1.35 V
Cross Talk	**350 mV**				
High-Frequency **Chip** **(>100 Mhz)** **ΔI = 20 Amps** **ΔT = 0.4 ns**	100 mV	Chip MCM	Thin-oxide 1.6 mm AVX C4 Ceramic (265nF)	164	713 nF CPchip 43 μF on MCM
Mid-Frequency **MCM** **(1 - 100 MHz)** **ΔI = 140 Amps** **ΔT = 2.5 ns**	65 mV	MCM Pu-Card	1.6 mm AVX C4 Ceramic (265nF) 1 μF Ceramic 10 μF Ceramic	164 1500 404	43 μF on MCM 1500 μF on Brd 4040 μF on Brd
Low-Frequency **Board** **(1 KHz - 1 MHz)** **ΔI = 235 Amps** **ΔT = 1 μs**	35 mV	Pu-Card	4700 μF Electrolytics	145	682 mF

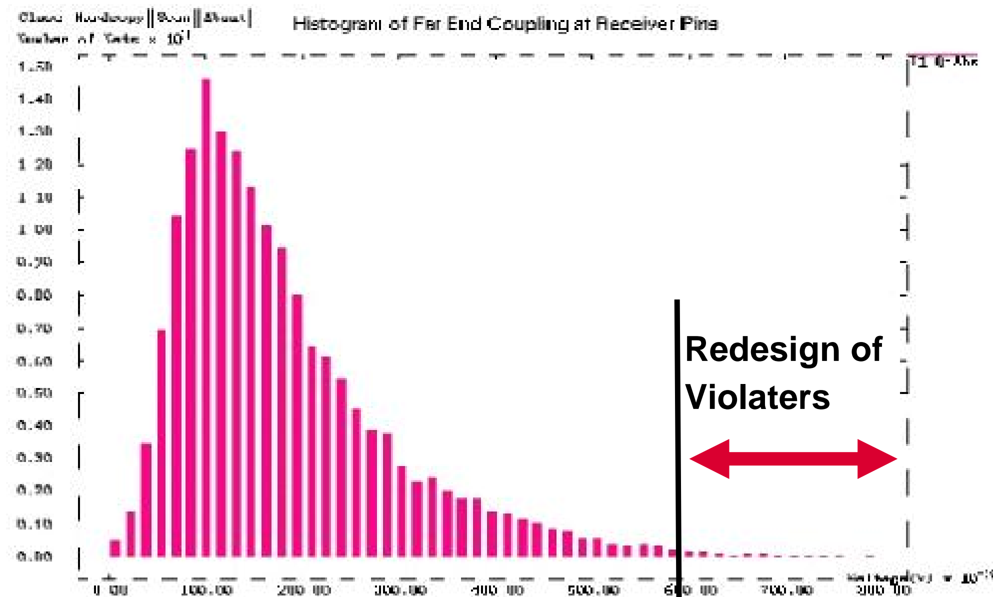


- noise budget consisting of lowfrequency, midfrequency and highfrequency noise
- midfrequency noise simulated by vendor tools

© 2008 IBM Corporation

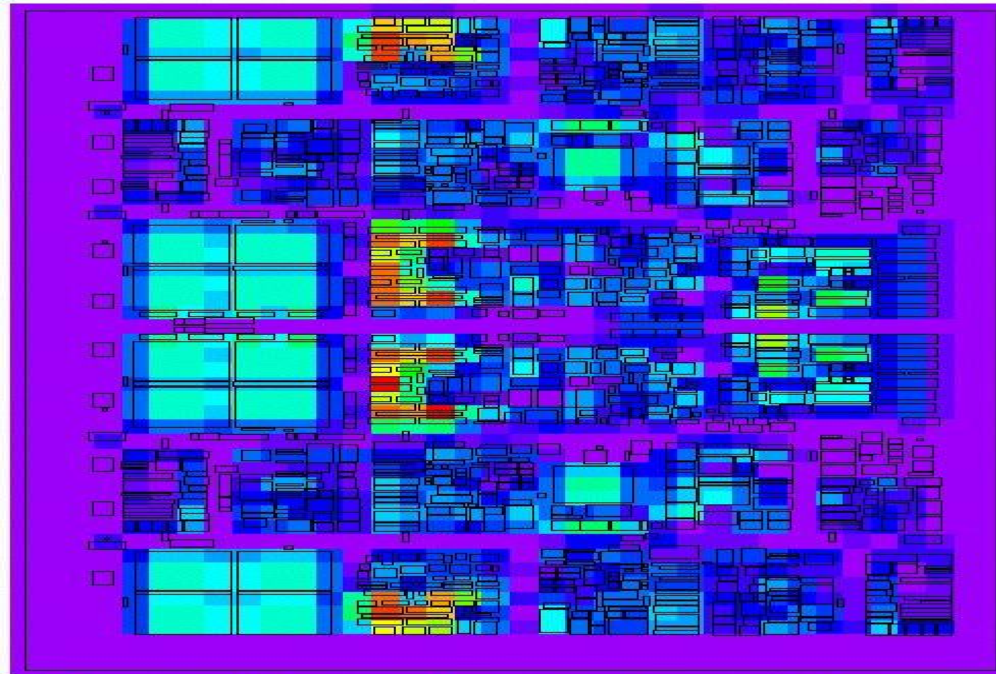
Post PD Timing and Noise Ananlysis

- system coupling noise run for each net (~12k) with in house tool
- about 5000 x faster than spice



© 2008 IBM Corporation

Highfrequency Analysis (Rapid)



- identification of hot spots on chip
- verify decoupling rules

© 2008 IBM Corporation

Summary and Conclusions:

- **Multi chip module technology enables architectures with huge bandwidths between processor and cache chips**
  - allows fully connected processor chips to all cache chips
  - not doable with today´s board technology
  - not doable with today´s organic technology

- **Bandwidth requirements will further increase when growing the number of processor cores in a system**
  - combined frequency and buswidth increase

- **High end server systems will continue to use high power chips**
  - system integration with larger number of cores on a chip
  - cooling techniques will enable > 200W chips
  - overall power saving on system level by integration

© 2008 IBM Corporation

Trademark Attribution Statement and Copyright

- IBM, the IBM logo, z9, z10, z Series, System z, System z9 and System z10 are registered trademarks of International Business Machines Corporation in the United States, other countries, or both.

- Other company, product, or service names may be trademarks or service marks of others.

- Copyright: Do not copy this lecture or any parts of it without the permission from the author.

© 2008 IBM Corporation