Key Challenges in High-End Server Interconnects

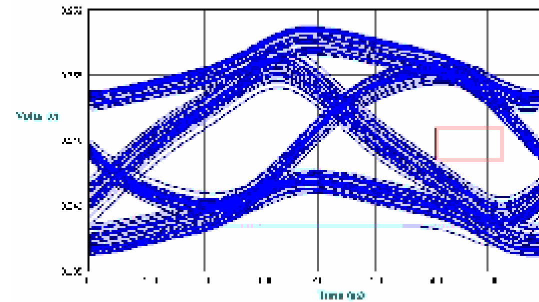
by Hubert Harrer

 © 2008 IBM Corporation

Packaging Key Challenges

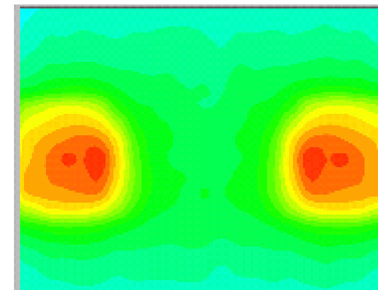
§**Bus Bandwidths**

- ►Large increase of bandwidths driven by multicore processors
  - –increasing frequency
  - –increasing bit lines



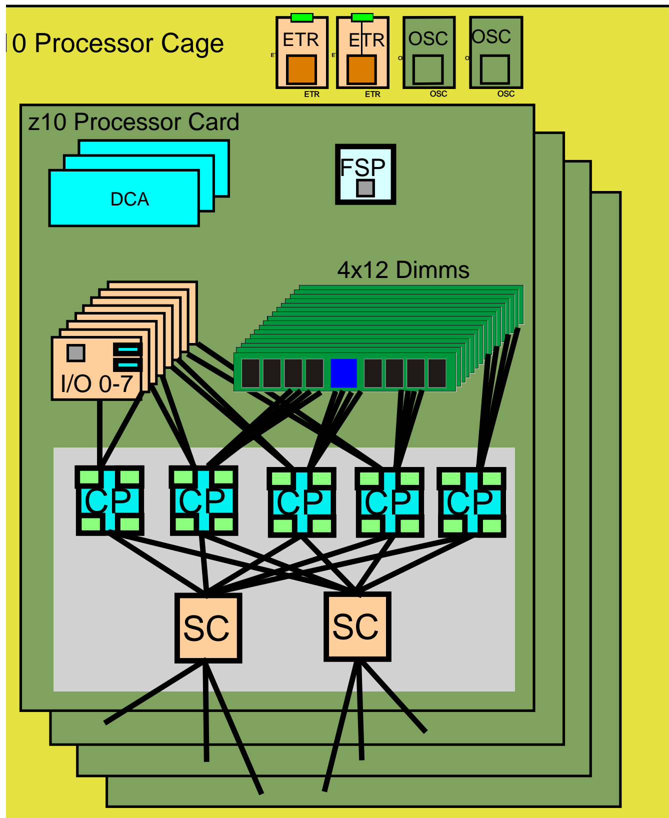
§**Power**

- ►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

Logical Structure of a System z High End Server

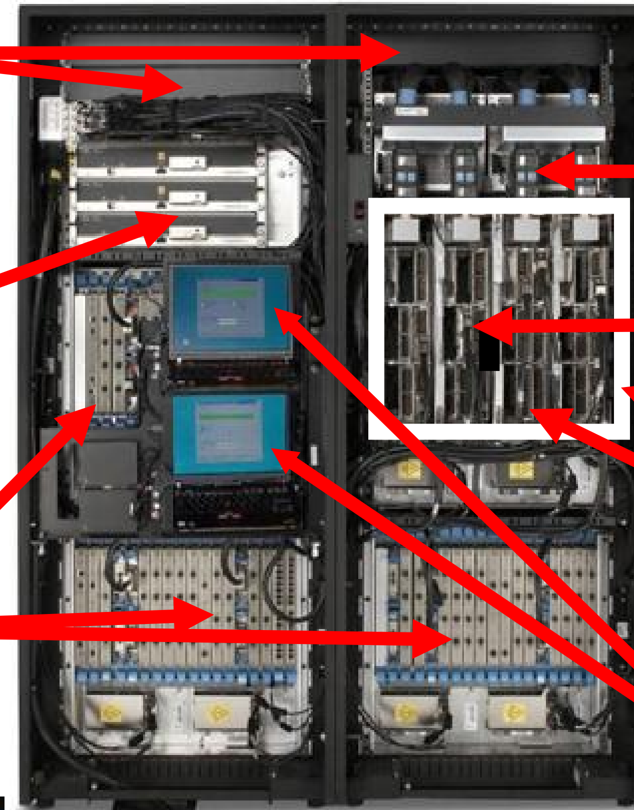


Internal Batteries (optional)

Power Supplies

3x I/O cages

Quick Connect Feature (optional)



Hybrid Cooling

Processor Books Memory

CEC Cage

STI cab

Support Elements

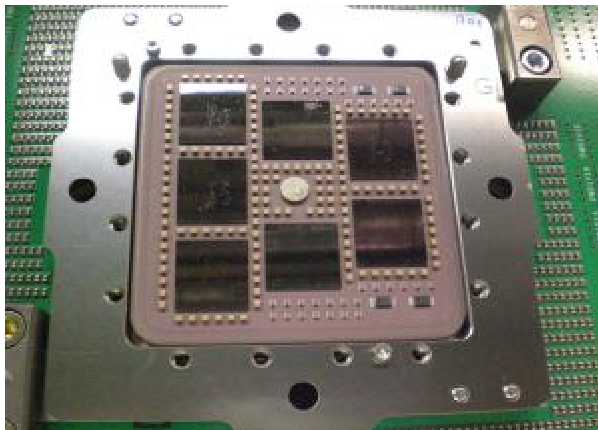
Front View

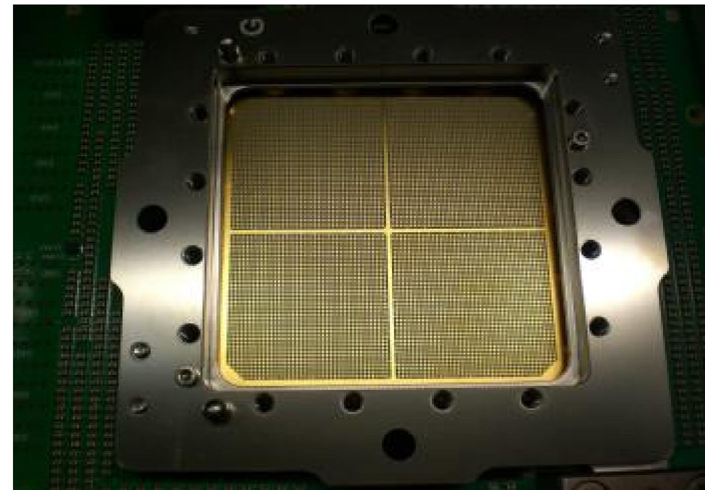
© 2008 IBM Corporation

:10-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

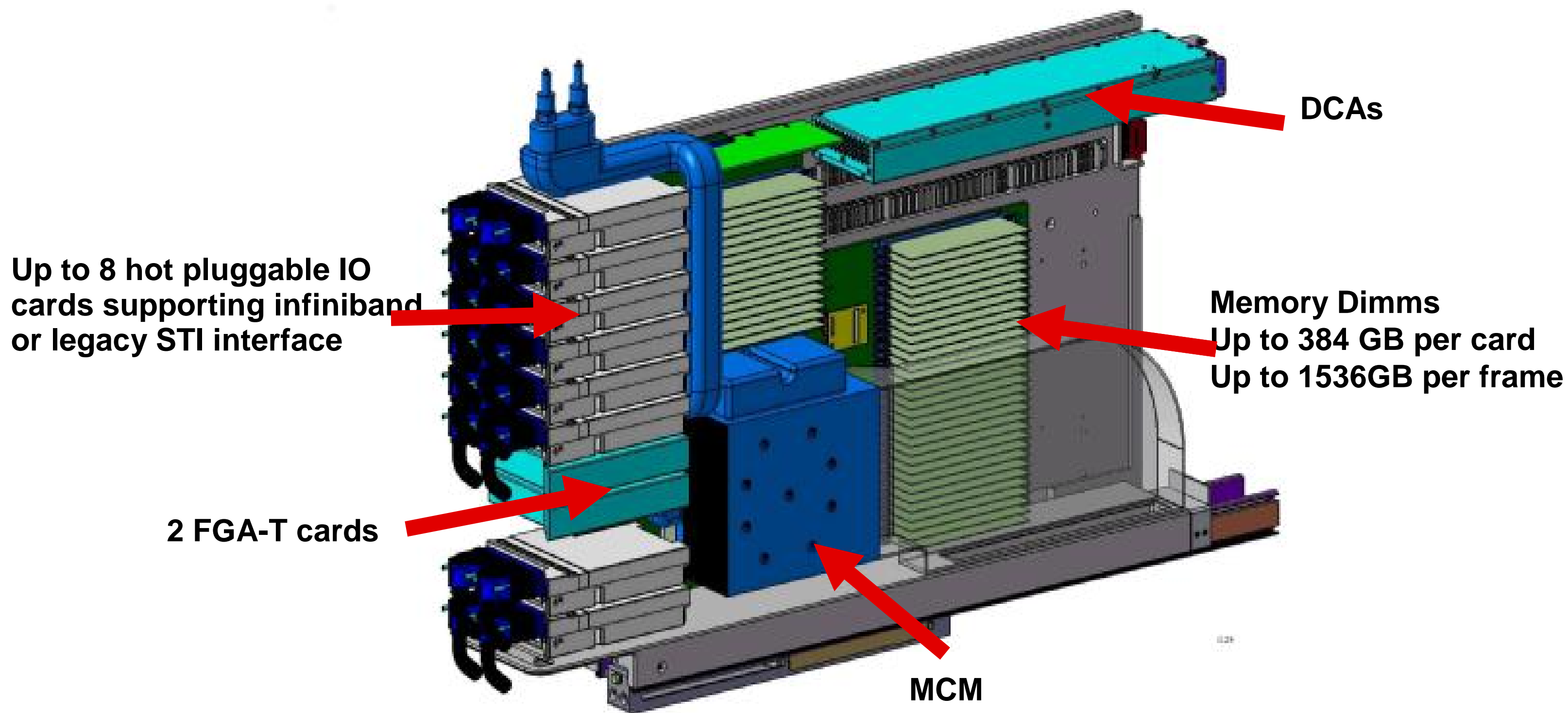


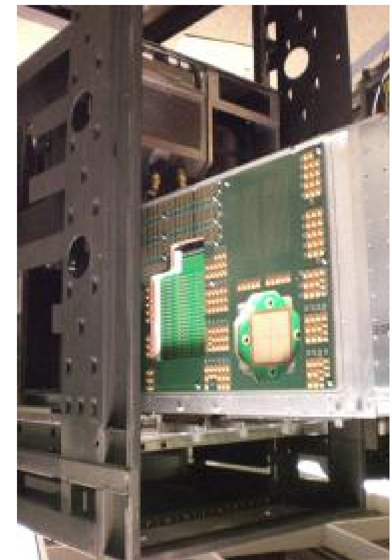


© 2008 IBM Corporation

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





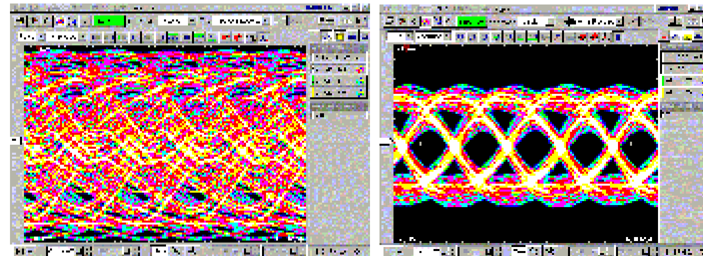
© 2008 IBM Corporation

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

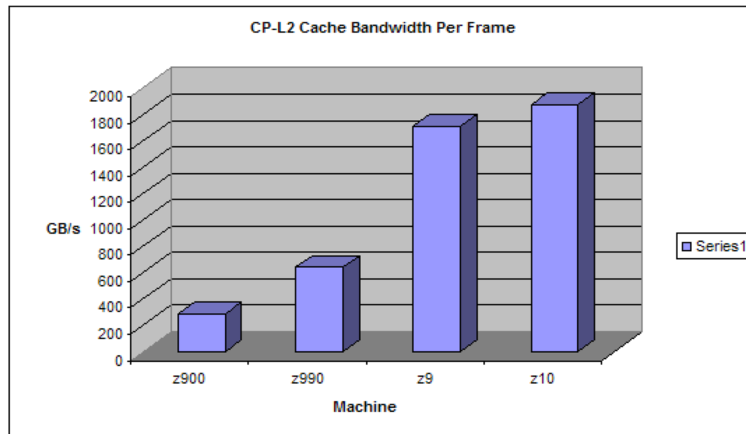
z9/z10 Timing and Bus Performance Comparison

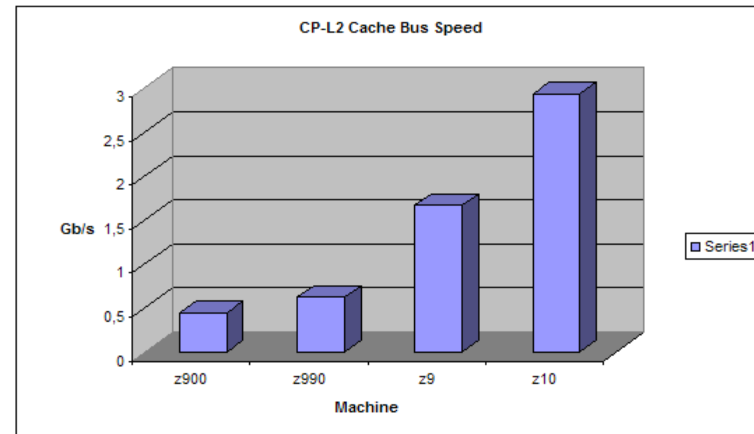
Name	Data Size B	Bit Rate Gb/s	Bandwidth GB/s 4 nodes	1st level Wire	2nd Level Wire	Comment
Cache Bus	256/160 x4	1.72/2.93	1764/1877	13 cm	n.a.	DDR source synch
Memory	64/48 x4	0.86/2.17	220/417	4 cm	20 in	DDR source synch
I/O Hub (MBA)	64/64 x4	0.86/2.2	220/563	3 cm	20 in	DDR source synch
Ring/Fabric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

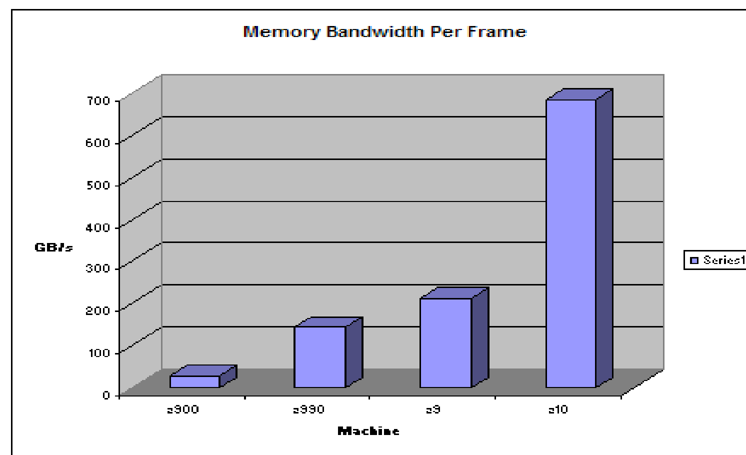
- wiring rule for each net over all components
- system timing run for each net (~12k) with in house tool
- single ended source terminated driver and diode clamped receiver for all nets besides STI

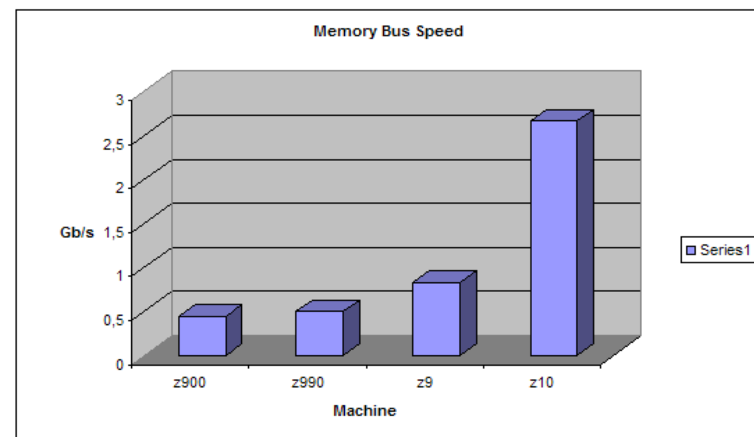
© 2008 IBM Corporation

Bandwidth Comparison









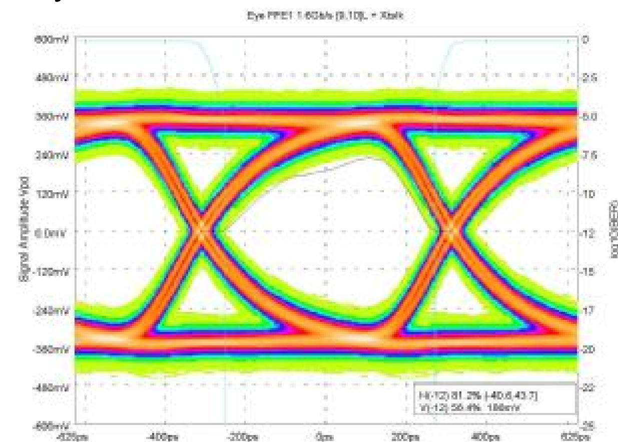
© 2008 IBM Corporation

Packaging Design Methodologies

§**Timing and Noise Methodology**

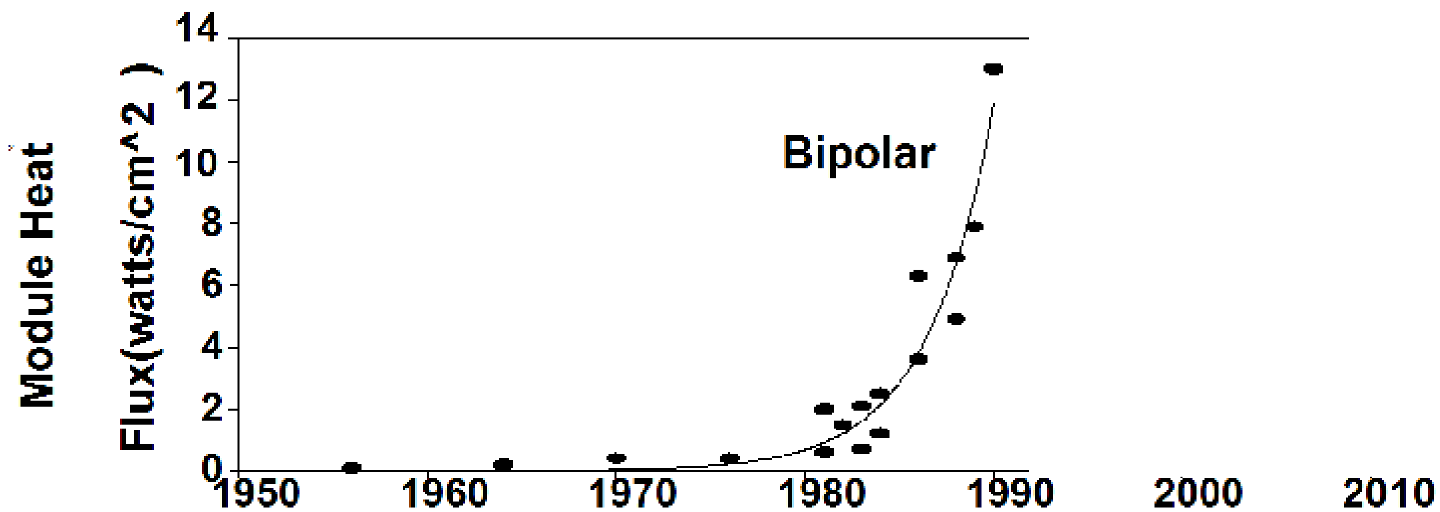
- ►**Pre PD:**
  - –create 3d models for all components (lines, vias, connectors)
  - –power spice simulation
    - Ÿuse length estimates from wiring analysis
    - Ÿuse coupling estimates from previous system experience
  - –HSSCDR (eye timing tool calculating bit errors probability) (IBM tool)
    - Ÿuses s-parameters for channel working in frequency domain
    - Ÿlinear model of driver
    - Ÿdriver jitter assumptions included

- ►**Post PD:**
  - –Fastline simulation for timing and noise (IBM Tool)
  - –all nets in the system are being simulated



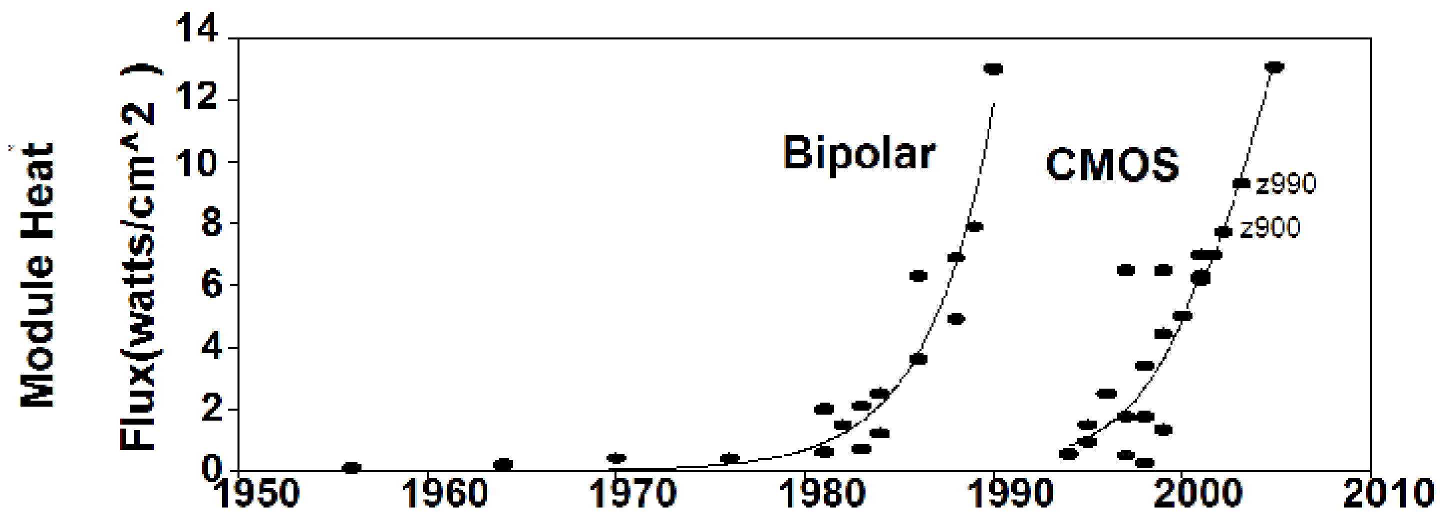
© 2008 IBM Corporation

Comparison of Bipolar and CMOS Power



© 2008 IBM Corporation

Comparison of Bipolar and CMOS Power



© 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

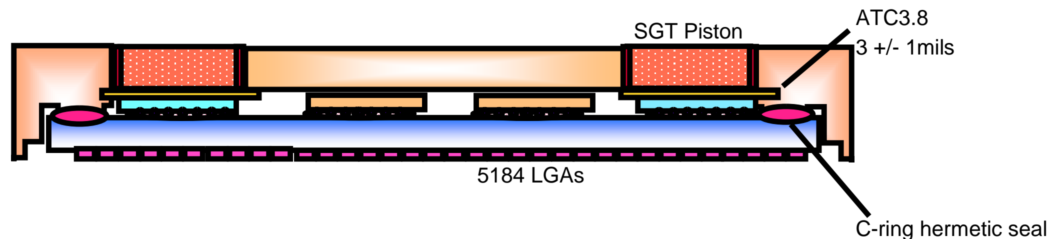
§**Power**

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

§**Temperature**

- ►45C chip junction

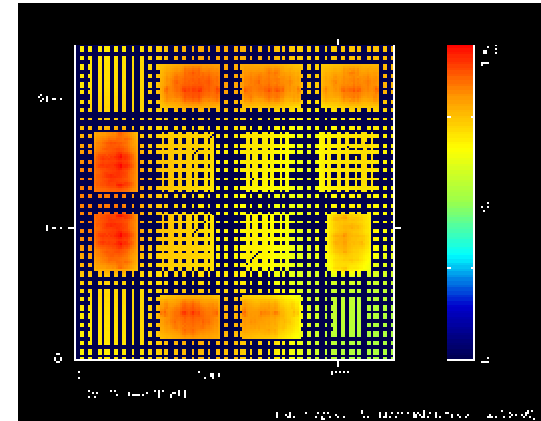
z9 MRU cooling
with aircooled backup mode

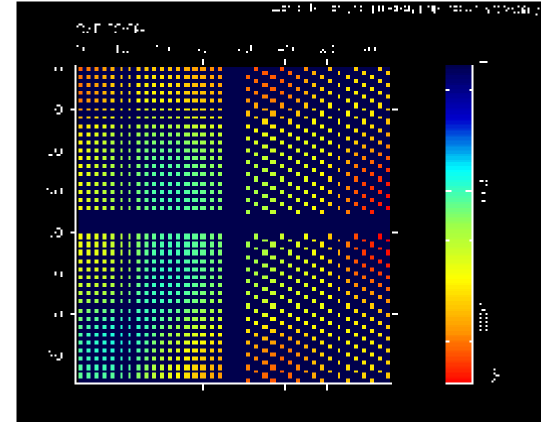


© 2008 IBM Corporation

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





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

© 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