The 3rd Generation of IBM's Elastic Interface on POWER6™

Daniel Dreps

Agenda

- Design Point Selection
- POWER6™ chip details
- Circuit approaches
- Testability and Hardware Measurements
- System design rules
- A System picture
- Conclusions

- **POWER6 & Z6 System Interconnect Requirements**
  - Shortest Net are on module Glass Ceramic hops
  - Longest board nets are 32 inches with 3 boards processor card to backplane to processor card
  - Longest cable net replaces the backplane with 2ft+ flex cable.
  - Backwards compatibility with EI-1 and E1-2
  - Up to 3.2 Gbit/wire was needed for performance for DDR2 (memory)
  - Must be easily portable between 65nm SOI and 90nm Bulk
  - Over 800 lanes coming off the processor is needed for system performance and scaling

- **POWER6 & Z6 System Interconnect Goals**
  - Area must be not more than 15% processor area.
  - Power is paramount and will limit the number of lanes in max configuration.
  - Leakage Power must be small for applications that wire out fractional I/O content. (Low-End , Mid-Range)
  - Minimize critical Analog circuit content.
  - Maximize Interface on chip diagnostics in area and power envelope
  - I/O cannot limit bring-up or time to market.

Requirement was to provide 3Gbit/wire (SE) or 6Gbit (DE) for performance

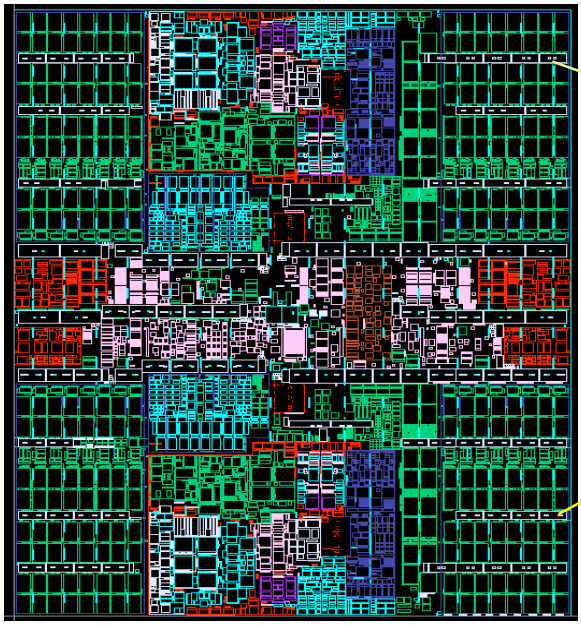
- **Single Ended considerations**
  - Can we control noise do to SSO, Xtalk, vref wander, via fields, module breakouts
  - Attenuation < -12dB on 32 inch 3 board backplane nets and flex cables
  - Backwards compatibility with EI-1 and E1-2
  - Use FR4(Flame Resistant 4) boards on all but P High End and Z6 systems
  - Signal through a SMT_DIMM connector onto an Integrated DIMM
  - Chip placement and area and power low enough .. solutions for both 65 nm SOI and 90 nm bulk
  - Be organic module tolerant.
  - Will need new connector , new High-End board material
  - Incremental Clocking improvements, PLL RJ (Random Jitter)

- **Differential considerations**
  - 2X on each wire
  - Long 32 inch 3 board backplane nets require Transmitter FFE3 (Feed Forward Equalization with 3 taps) and Receiver DFE5 (Decision Feedback Eqaulization with 5 taps) which is large area and large power
  - Many PLL's and special CML (Current Mode Logic) clock distributions required
  - Special board steps required .. Back drilling under backplane connectors
  - Chip placement is constrained I/O under C4 bumps to meet return loss.

- Design Point Selection
- POWER6™ chip details
- Circuit approaches
- Testability and Hardware Measurements
- System design rules
- A System picture
- Conclusions

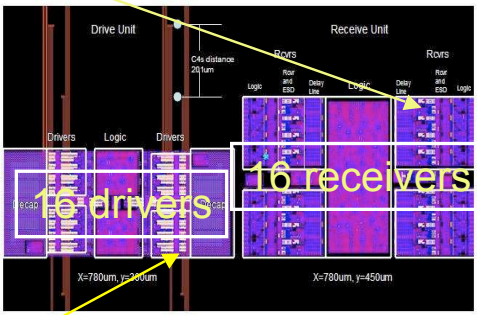
POWER6™ has 811 EI-3 lanes flexible placement not perimeter only

Power6™ Microprocessor



9 Signal Cols 12 Signal Cols

16 pack units magnified

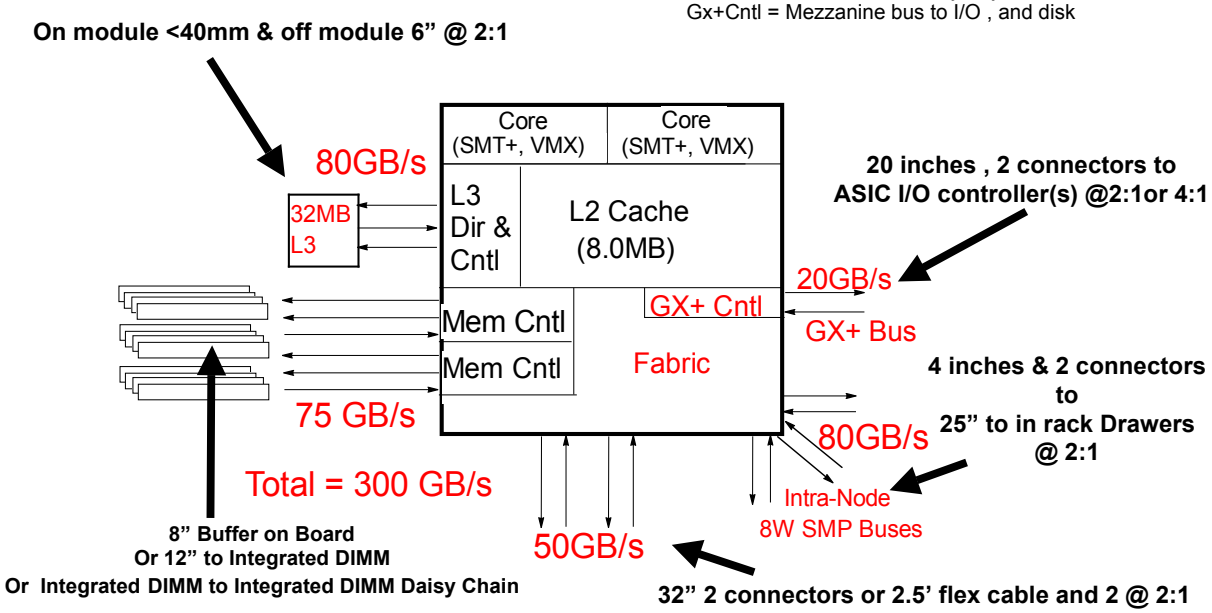


I/O in 16 packs “area array”
Most white boxes have 16 drivers or 16 receivers. An 8Byte chip to chip bus has 10 white boxes 5D , 5R. Each white box has 2 C4 signal columns and 2 C4 power columns.

POWER6™ Scales Chip Bandwidth with Core Performance

300 GB/s total IO Bandwidth

SMT+ = simultaneous multi- threading
VMX = vector multimedia extension
L3 Dir & cntl= 3rd level cache directory & controller
Dir = L3 directory
Mem Cntl = on board memory Async control
Gx+Cntl = Mezzanine bus to I/O , and disk

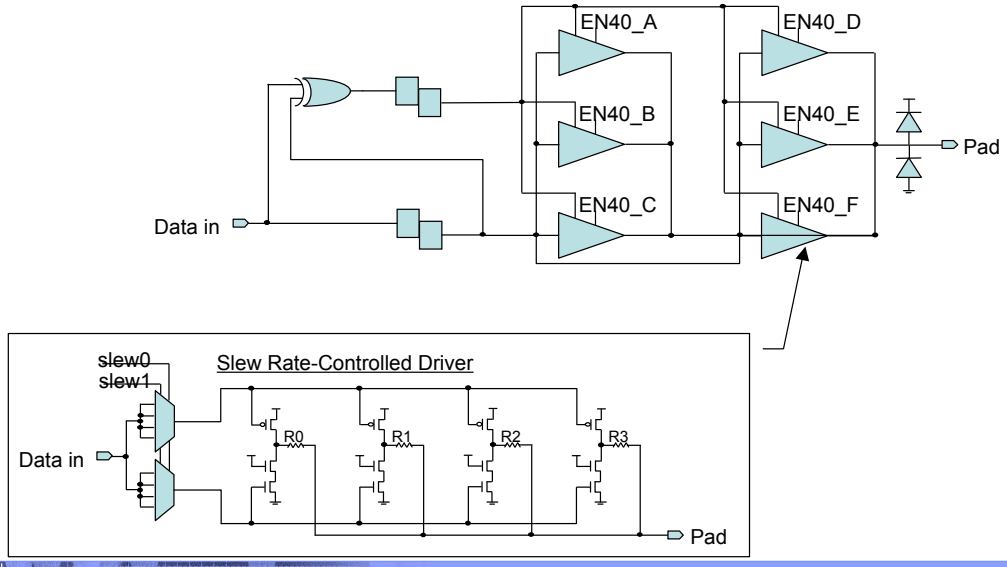


- Design Point Selection
- POWER6™ chip details
- Circuit approaches
- Testability and Hardware Measurements
- System design rules
- A System picture
- Conclusions

- **Circuit diagrams**
  - Output Driver Diagram and Mode Pin Requirements
  - Receiver Diagram and Phase Adjustment
  - Analog Delay Line
  - Lane Sparing Diagram
  - Layout Blow-up of Driver and Receiver Packs

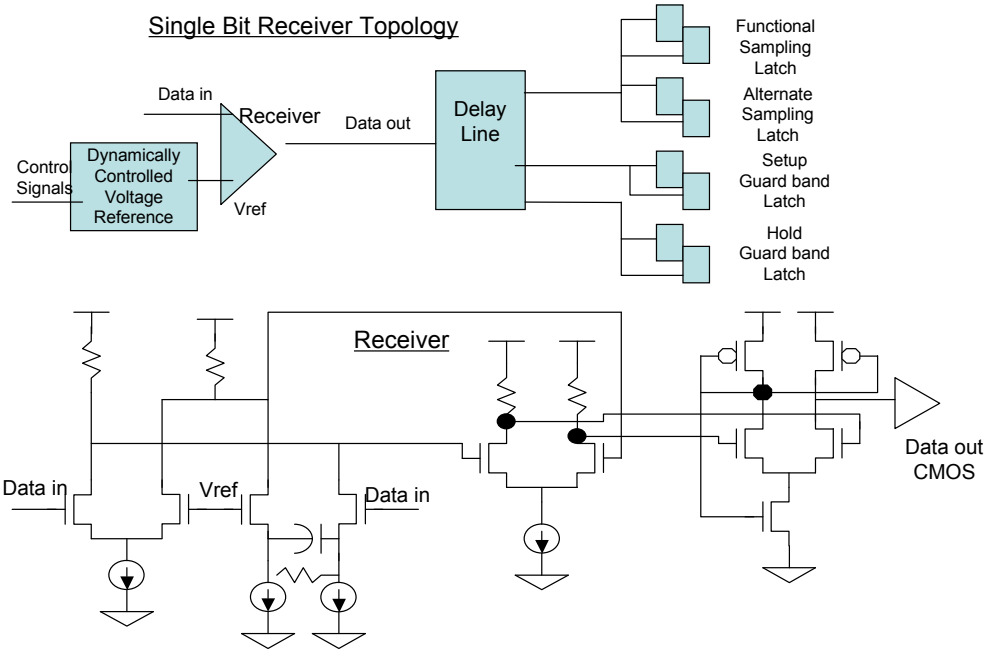
Key Silicon design element EI-3 FFE-2 Driver Topology

- SE CMOS design with FFE2 (Feed Forward Equalization with 2 taps).
- A dedicated I/O rail is used and is nominally set to 1.2V
- The first tap can be programmed to 6.6, 8, 10, 13.3, 20 and 40 ohms.
- The second tap is 40 ohms.
- The driver also has 4 levels of slew rate control to minimize SSO (Simultaneous Switching Outputs)
- The driver's impedance matching is done without any external resistors and is achieved with a precision poly silicon resistor that has +/-10% tolerance.
- ESD (Electrostatic Discharge Devices) were designed to minimize wiring but also allow for flexible C4 to pad wiring. We can span 1100 Microns on Last Metal Pad Transfer Wiring.



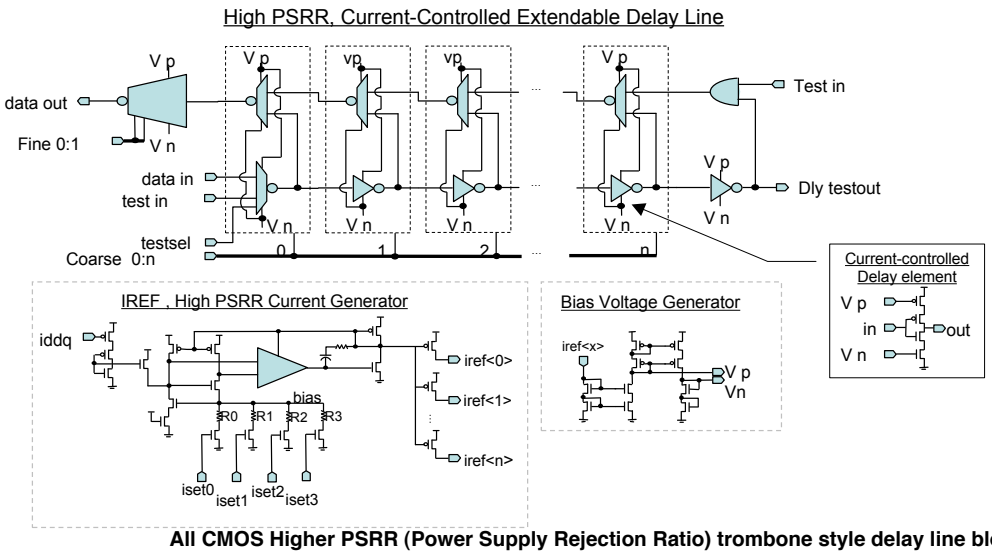
Key Silicon Design Element EI-3 Rcvr Topology

- Nfet CML front end with peaking vector
- Learned Vref per bit bandwidth optimized, steps of +/- 10 mV for 300 mV range
- Flops are all static LSSD(Level Sensitive Scan Devices)



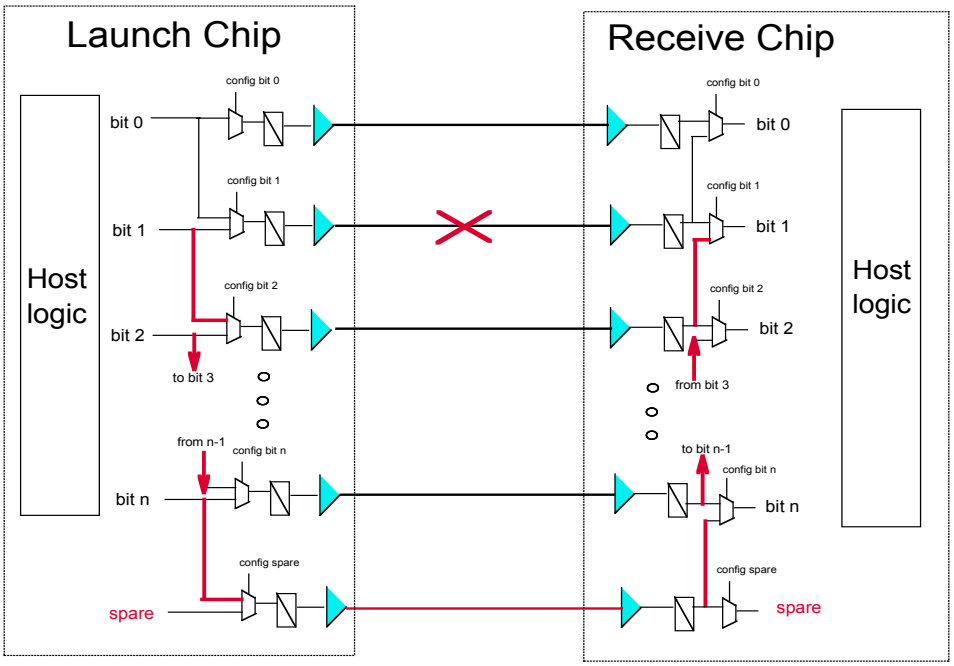
Key Silicon Design Element EI-3 Rcvr Delay Line

- Each per bit delay line as 4 taps, Set-up (S), Hold (H), Functional (F), Alternate (A)
- Set-up marks the leading edge of open eye, Hold marks the trailing edge
- Functional and Alternate Ping-Pong under algorithm control for clocking and temperature drift.
- Using these taps we can calculate the eye opening per bit real time



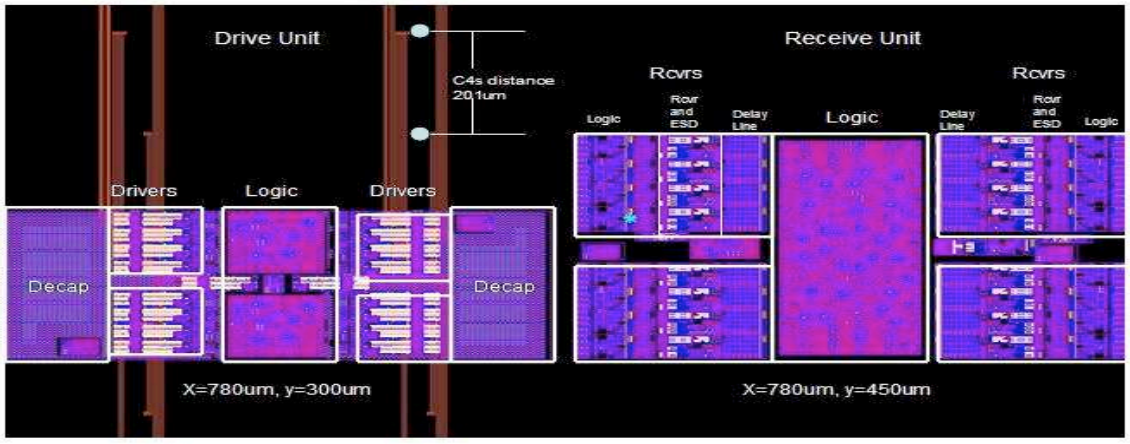
All CMOS Higher PSRR (Power Supply Rejection Ratio) trombone style delay line block

Logical Diagram of electronic repair



Blow-up of area of a 16 bits of receiver and 16 bits of driver.

- 8 Byte BUS is assembled with 10 packs of 16 bits, 5 driver packs and 5 receiver packs
- A Source Synchronous Differential DDR (Double Data Rate) clock exists for each 16bit.
- Each BUS has a driver state machine controller and a receiver state machine controller
- FLY Wires on LAST METAL reach up to 1100 microns flying over L2 (second level) cache.
- Area per bit of TX =.014 sq mm Area per bit of RX =.021 sq mm



- Design Point Selection
- POWER6™ chip details
- Circuit approaches
- Testability and Hardware Measurements
- System design rules
- A System picture
- Conclusions

EI-3 has a Virtual Oscilloscope for every bit.

- Running BIST , sweeping the test pattern we can determine the passing delay steps represented by “dashes”
- We can also plot the A H F S to see if the A and F taps show the same EYE position since in drift mode we maybe wrapping around the delay line since we have 360 degree phase tracking with a delay line approach.
- The registers can be accessed real time for debugging if a Correctable Error were to occur.

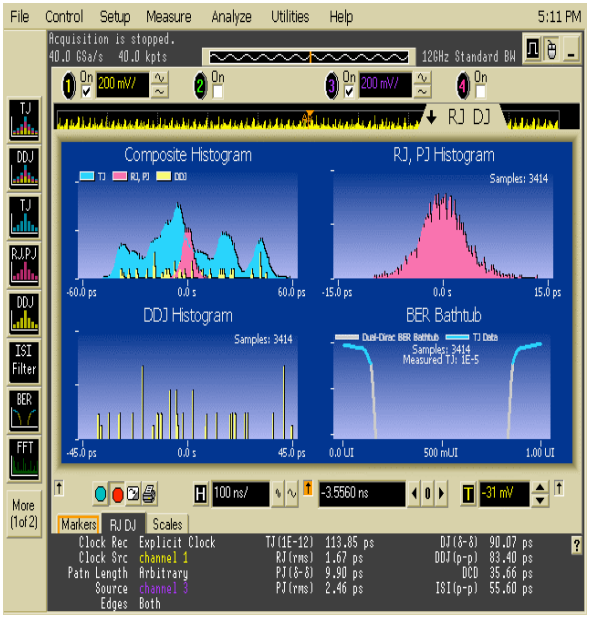
- **Delay tracking measured internally for 1 bit lane**
- **The “A (Alternate)” and “F (funcunctional)” paths have GOOD tracking within 2 steps**

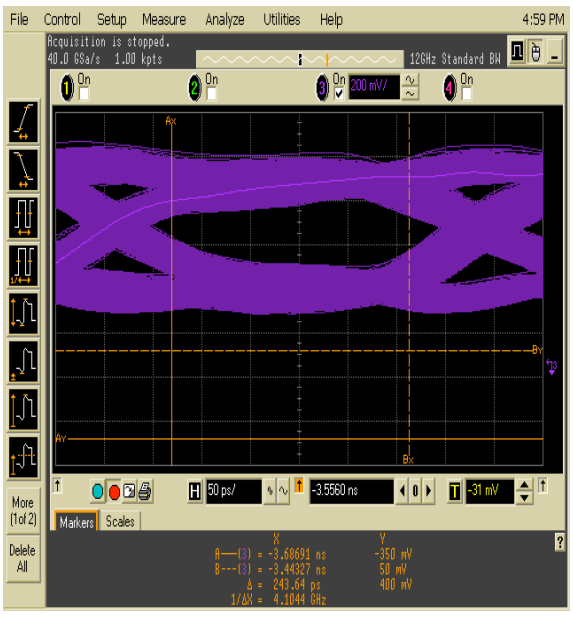
```
.       rcv  rcv  rcv  rcv rcv          .   .    .   sel  ....-1.........0.........1.........2.........3.........4.......
sampler chip cage node pos bus          grp pack bit edge 543210987654321x12345678901234567890123456789012345678901234567
------- ---- ---- ---- --- ---          --- ---- --- ---- ---------------------------------------------------------------
flags   p6    0    0    0  PRT_EI_FBC_X  1   0     0   1   11111...................A......H.............F.............S....
alt     p6    0    0    0  PRT_EI_FBC_X  1   0     0   0   00000..........XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
alt     p6    0    0    0  PRT_EI_FBC_X  1   0     0   1   11111..........XXXXXXXXXXXXXXXXX----------------------------XXXXX
alt     p6    0    0    0  PRT_EI_FBC_X  1   0     0   2   22222..........-------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
alt     p6    0    0    0  PRT_EI_FBC_X  1   0     0   3   33333..........XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
func    p6    0    0    0  PRT_EI_FBC_X  1   0     0   0   00000..........XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
func    p6    0    0    0  PRT_EI_FBC_X  1   0     0   1   11111..........XXXXXXXXXXXXXXXXX----------------------------XXXXXX
func    p6    0    0    0  PRT_EI_FBC_X  1   0     0   2   22222..........-------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
func    p6    0    0    0  PRT_EI_FBC_X  1   0     0   3   33333..........XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

Passing delay settings window.

Measurements

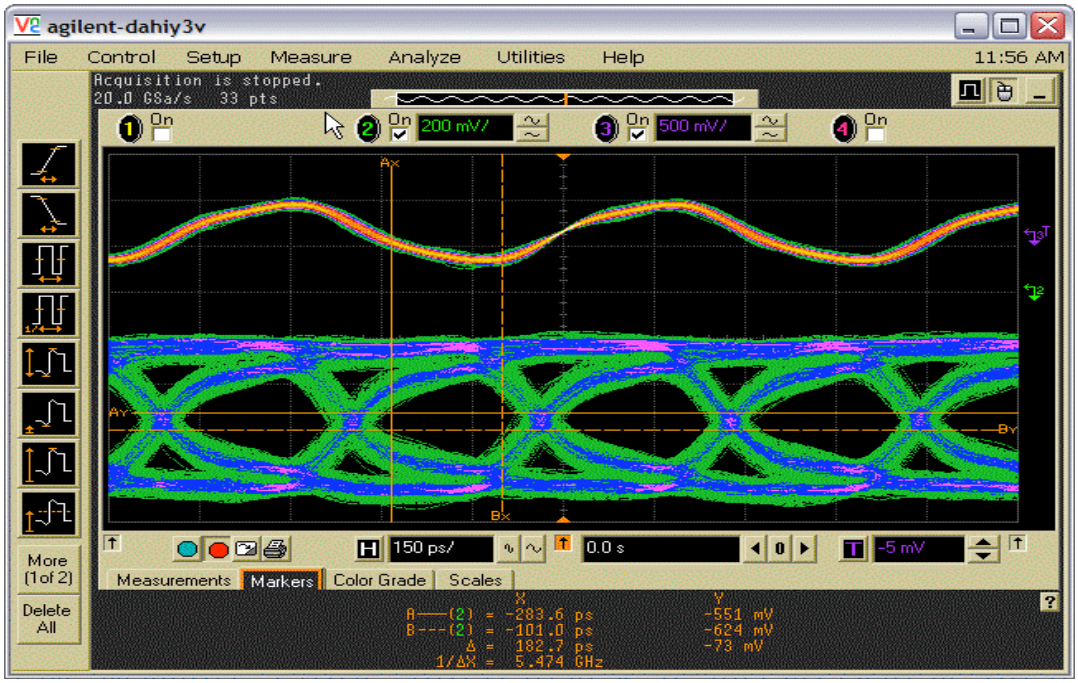
- Extracted Random Jitter(RJ), <2 ps using with launch to capture with +/- 1UI
- EYE opening at the BSM of P6 for 40 cm , 3 board connection
- EYE opening at +/- 50 mV was 68% for this case
- Digital EYE extraction inside chip agrees within 15%





3.2Gbits EYE (MEMORY CHANNEL_ P6 driving Integrated DIMM through SMT-DIMM connector

Eye opening at +/- 50 mV @ 3.2Gb/s 182ps/312ps = 75 % , 10 minutes Random Data Pattern (BIST)



- **Design Point Selection**
- **POWER6™ chip details**
- **Circuit approaches**
- **Testability and Hardware Measurements**
- **System design rules**
- **A System picture**
- **Conclusions**

Key design aspects for system success, the packaging aspects required:

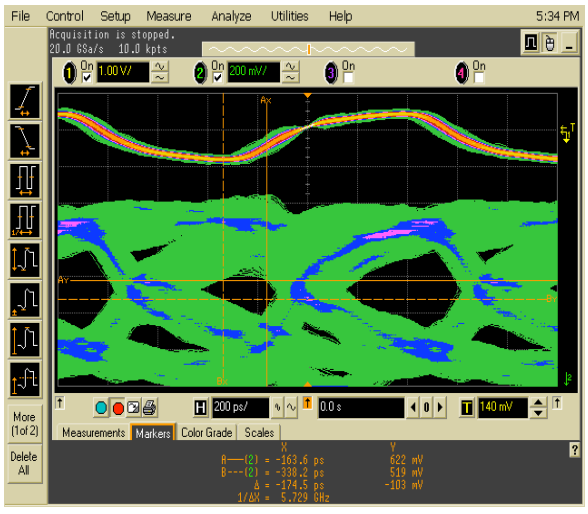
- Return path design
- DIMM connector
- LGA
- Cable and Backplane Connector
- Board wiring rules exploiting Source Synchronous
- On chip clocking and distribution noise
- Scrambling
- Power supply noise budgets

Return Path Design

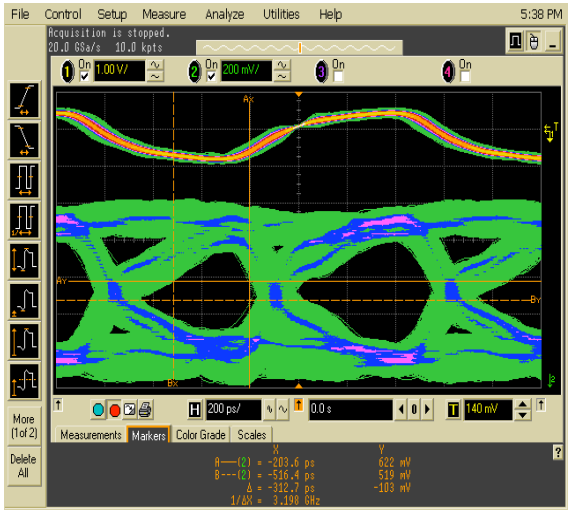
- Integrated DIMM ASIC is on Ball Grid Array (BGA) with 0.8 mm pitch
- Multi chip modules with high power are compression connections using “cinch” connectors using dual compression retention
- Signal to Power ratios (S/P) were determined using 3D modeling. 2:1 S/P for receivers and 1:1 for drivers
- Engineering of anti pads and drill sizes were required to insure no impedance dipped below 40 ohms.
- 3D modeling via fields and connectors
- Grouping of Drivers and Receivers is required to isolate Near End (NE) to Far End (FE) coupling. No Duplex layout rules are allowed.

Measurement of Bad Return Path Isolation

DIMM S/P (Signal to Power) ratio was too aggressive on first DIMM hardware. Bad bit running our test pattern measurement compared to running every other lane inverted which is results in a parallel coded measurement, we have a test mode to invert our scramble pattern every adjacent bit to see if the return paths and cross talk change dramatically. This was a basic packaging diagnostic we used. We would measure EYE widths in both modes.



Bad bit in functional mode

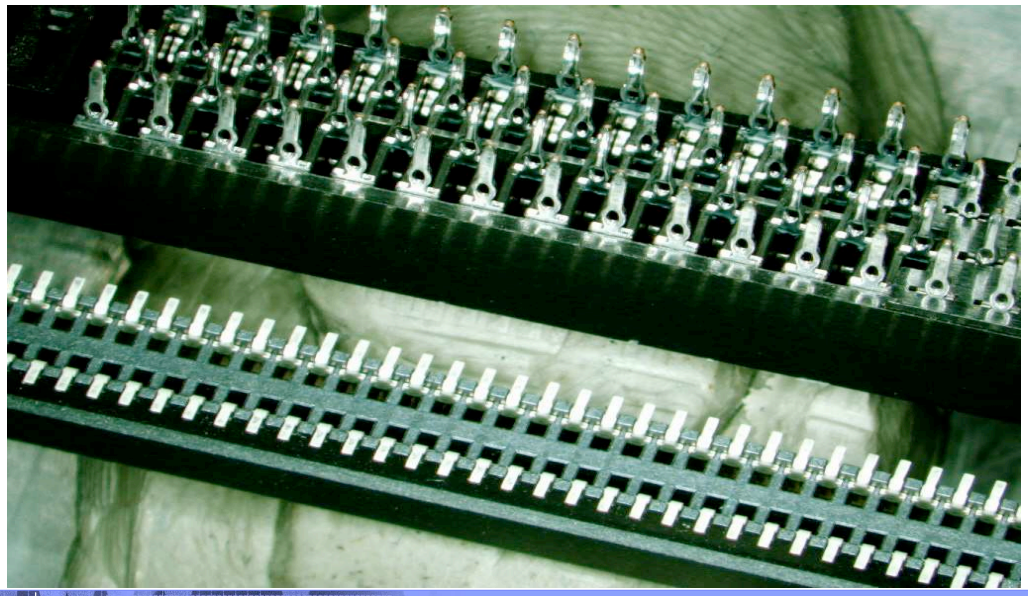


Bad bit with data parallel coded

SMT DIMM Connector

Used Gull Wing Surface Mount (SMT) instead of Plated Through Hole (PTH)

- Reduced crosstalk
- PTH finish (26 mil) was too large
- Large PTH also lowers the impedance



LGA (Land Grid Array) Connector

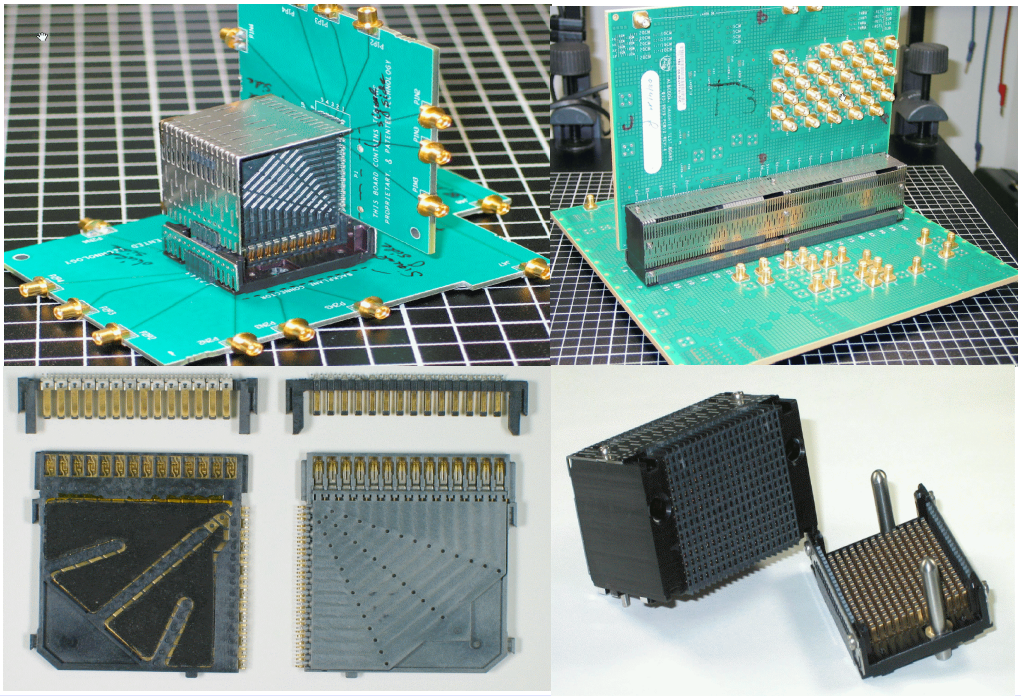
Fuzz button sockets used, as we have on POWER4 and POWER5 systems, the S/P (Signal to Power) Ratio we used was ~2:1 in receiver pin out areas and 1:1 in driver pin out areas. The 1:1 can be relaxed. To reduce large Driver near end noise propagating to Receivers, we also did not allow Driver and Receiver interdigitation, or bundling of Driver groups and Receiver groups.

Approx 0.8 mm height

CINCH WIRE BUTTON

Backplane Connector

The old legacy VHDM connector limited our bus speed to 2.0 Gbit for long 2 board connections. We later were able to lower driver dv/dt and adjust peaking to get VHDM and flex to run at 2:1 speeds. We use the Ventura connector for 2.5Gigabit signals and above on all backplane connectors.



Printed Circuit Board Wiring Rules

- FR4 4 mil one ounce 5 mil space on Blades to Midrange servers
- Megtron_6 boards on mainframes and P high end servers where total trace is up to 32”
- All bits is a clock group are wired end to end at +/- 1 UI (Unit Interval) of skew, we achieved 200 ps
- Limit thru loss (S21) to -12 dB at the fundamental
- Access crosstalk with worse case coupled design rules, allow ~ 50 ps for aggressors
- Organic module escape < 3 cm using 20 micron wires.
- No wires on organic over or near degassing holes
- Used fixed via designs to control impedance and cross talk under modules and for connectors. No changes to barrels, drill sizes or anti pads allowed.
- All nets are point to point uni-directional, BIDI for chip test only
- No transition vias, vias only under modules and connectors.

Clocking, Coding and Power Supply

Clocking

- Source Synchronous Clocking, a differential clock per 16 to 24 bits, CLOCK not STROBE
- Bus clock is DDR using both edges for sampling
- PLL is a ring type VCO (Voltage Controller Oscillator) with an RJ < 2ps over a +/- 2UI interval. LC tank not needed.
- Clock distribution is ALL CMOS with attention to placing as much delay in wire as possible.
- Clock tree latency of buffers optimized, LCR modeling required.
- All latches and local buffers are static CMOS.

Coding

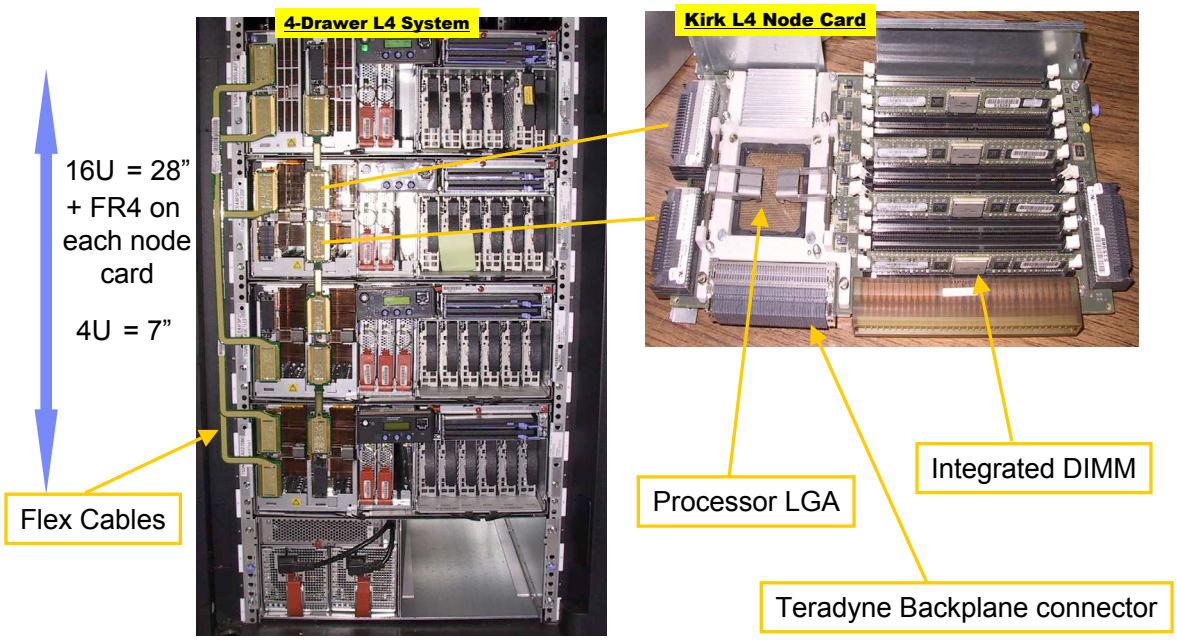
- Scrambling required on 65nm SOI parts to keep body voltage stable in delay line.
- On short links were 15% UI margin exist scrambling can be turned off for power savings.

Power Supply

- On processors allowable supply variation = +/- 10%, +/- 5% on 90 nm ASICS

- **Design Point Selection**
- **POWER6™ chip details**
- **Circuit approaches**
- **Testability and Hardware Measurements**
- **System design rules**
- **A System picture**
- **Conclusions**

Midrange System picture showing Flex Cabling and the processor node card



Conclusions

- **EI-3 links are highly flexible**
  - 32mm on module nets to 32” 3 boards & 2ft+ cables
  - Backwards compatible to EI-1 & EI-2
- **EI-3 link are very high frequency**
  - 3.2Gbit Single Ended
- **EI-3 Links are very high density**
  - < 32mm2 for 811 lanes on a 65nm SOI processor
- **EI-3 Links are very low power**
  - 32W for 811 lanes @ >2.5 Gbit
- **These constraints could not have all been met without optimization of all interconnect elements in the system**
- **Time to market goals could not have been met without the extensive investment in built-in debug features**

Acknowledgements

- Rob Reese , Pete Thomsen, Mike Spear , Racheal Eberly, John Gullickson for logic design and verification.
- Frank Ferraiolo for leadership
- Glen Weidemeier, John Schiff ,Hector Saenz ,Mike Sperling ,Bao Troung and Seongwon Kim for circuit design and timing and special analog design rules
- George Chu and Tommy Tidwell for layout and unit build
- Anand Haridass and Dulce Altabella and Mark Ritter and Lei Shan for Signal Integrity and SI design rules
- Humberto Casal and Gary Peterson and Megan Nguyen for Lab bring-up debug and diags and link mode hardware optimizations.

