

New IBM Components for HPCx

HPCx Annual Seminar
December, 2003

Charles Grassl
IBM



Agenda

- **POWER4+**
- **POWER5**
- **High Performance Switch**



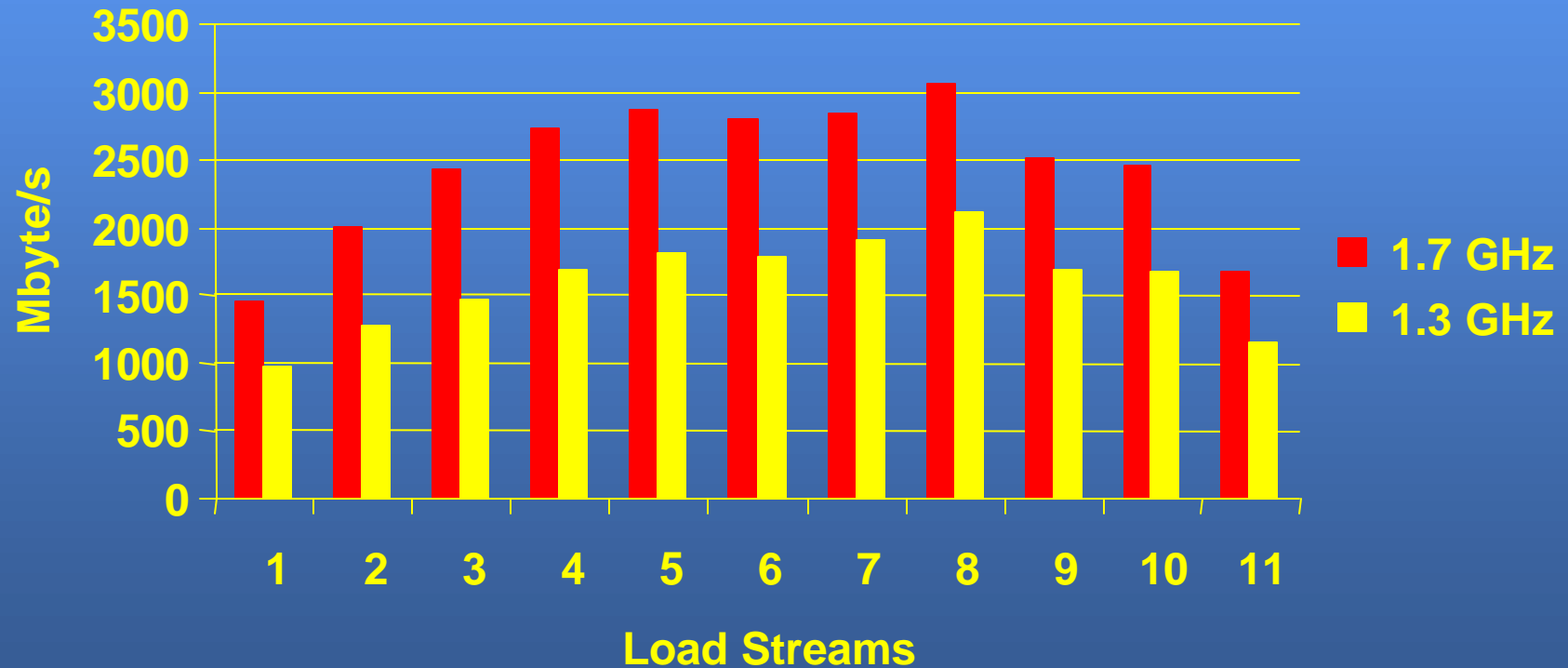
POWER4+

- **Increased clock rate**
 - 1.3 GHz → 1.7 GHz (+30%)
- **Improved L3 Cache**
 - **Buffers “fixed”**
 - Reduced contention
 - Improved stores
- **Increased L2 cache size**
 - 1.44 Mbyte → 1.5 Mbyte
 - **“Fully” 8 way set associative**



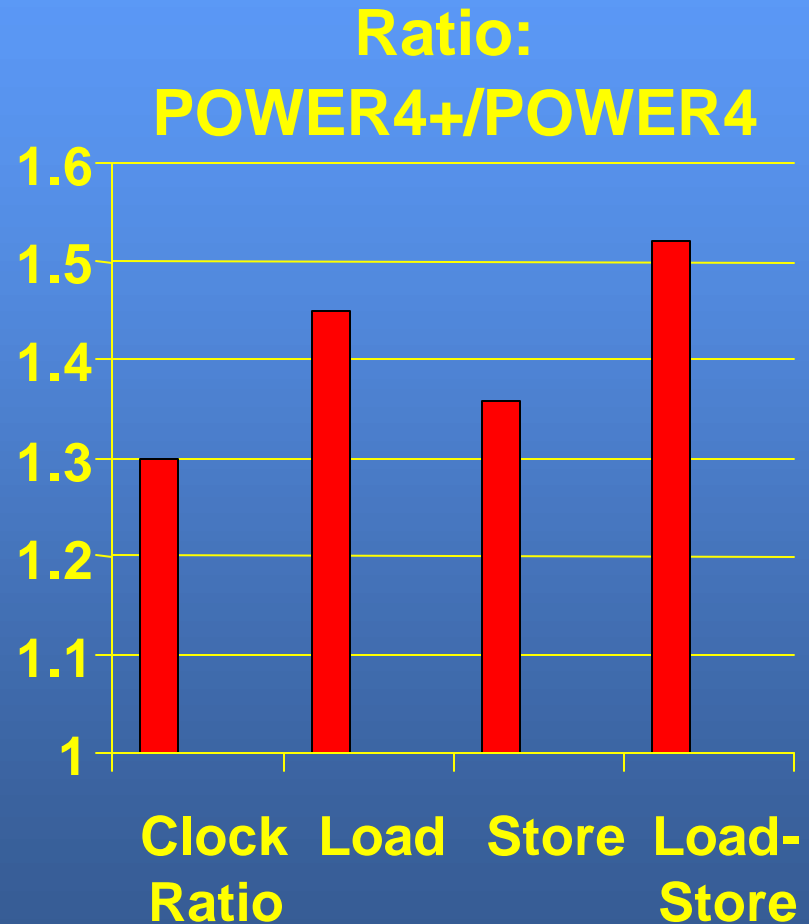
POWER4+: Memory Bandwidth

Load - Store Operations



Bandwidth Improvement

- **Load: 45%**
- **Store: 35%**
- **Load – Store: 50%**



1.0	1.1	1.2	1.3	1.4	1.5	1.6
1.7	1.8	1.9	2.0	2.1	2.2	2.3
2.4	2.5	2.6	2.7	2.8	2.9	3.0
3.1	3.2	3.3	3.4	3.5	3.6	3.7
3.8	3.9	4.0	4.1	4.2	4.3	4.4
4.5	4.6	4.7	4.8	4.9	5.0	5.1
5.2	5.3	5.4	5.5	5.6	5.7	5.8
5.9	6.0	6.1	6.2	6.3	6.4	6.5
6.6	6.7	6.8	6.9	7.0	7.1	7.2
7.3	7.4	7.5	7.6	7.7	7.8	7.9
8.0	8.1	8.2	8.3	8.4	8.5	8.6
8.7	8.8	8.9	9.0	9.1	9.2	9.3
9.4	9.5	9.6	9.7	9.8	9.9	10.0

POWER4+ Physical

- ~Same number of transistors
- **Smaller die**
 - 400 mm² → 380 mm²
 - 170 → 130 nanometer line widths
- **Lower power consumption**

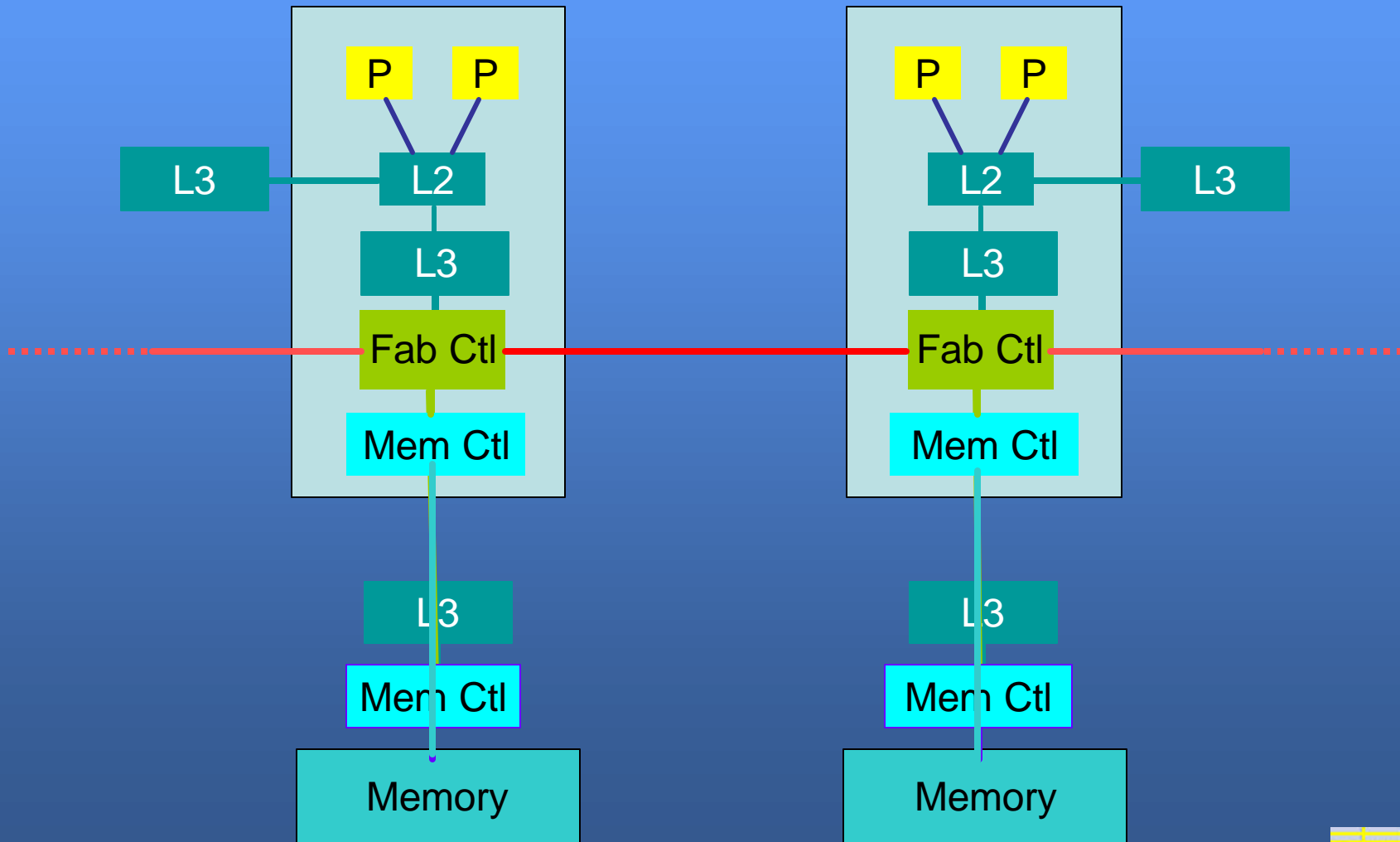


POWER5: Objectives

- **POWER4 base:**
 - Binary and structural compatibility
 - Enhance and extend SMP scalability
- **Performance**
- **Power efficient design**
- **Reliability, availability, serviceability attributes**

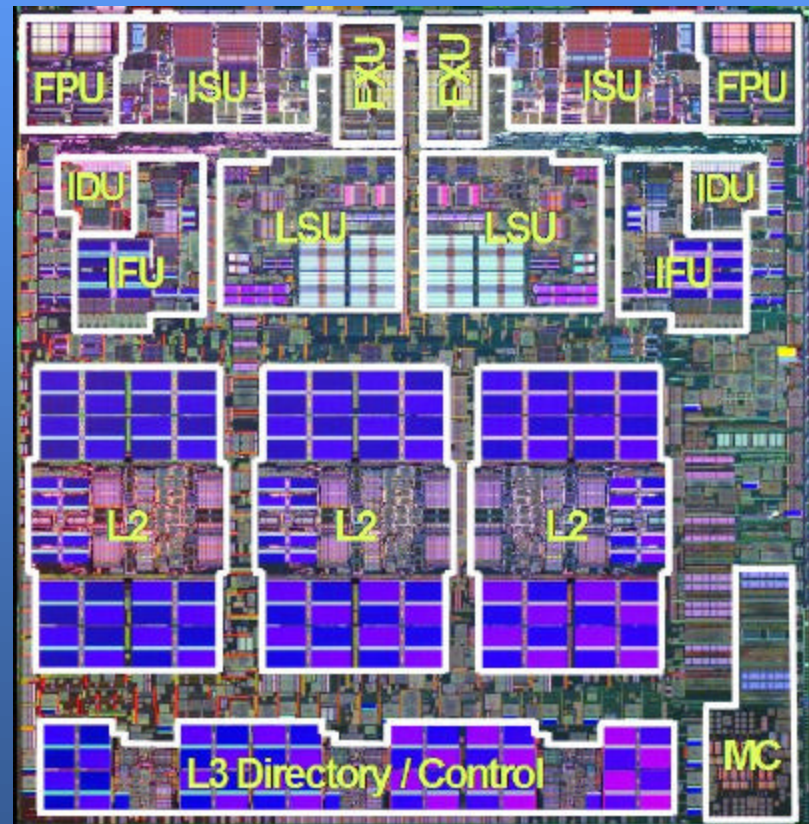


Modifications to POWER4 System Structure



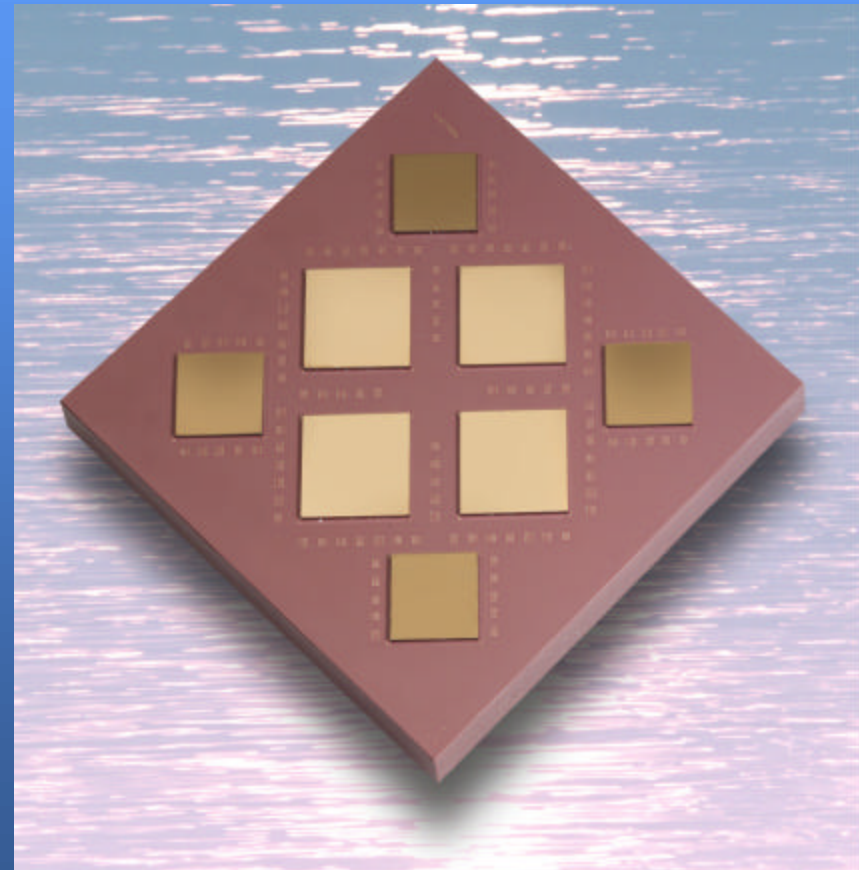
POWER5 Chip

- **IBM CMOS 130nm**
 - Copper and SOI
 - 8 layers of metal
- **Chip**
 - 389 mm²
 - 276 M transistors
 - I/Os: 2313 signal, 3057 power



POWER5 Multi-Chip Module

- 95mm × 95mm
- Four POWER5 chips
- Four cache chips
- 4,491 signal I/Os
- 89 layers of metal

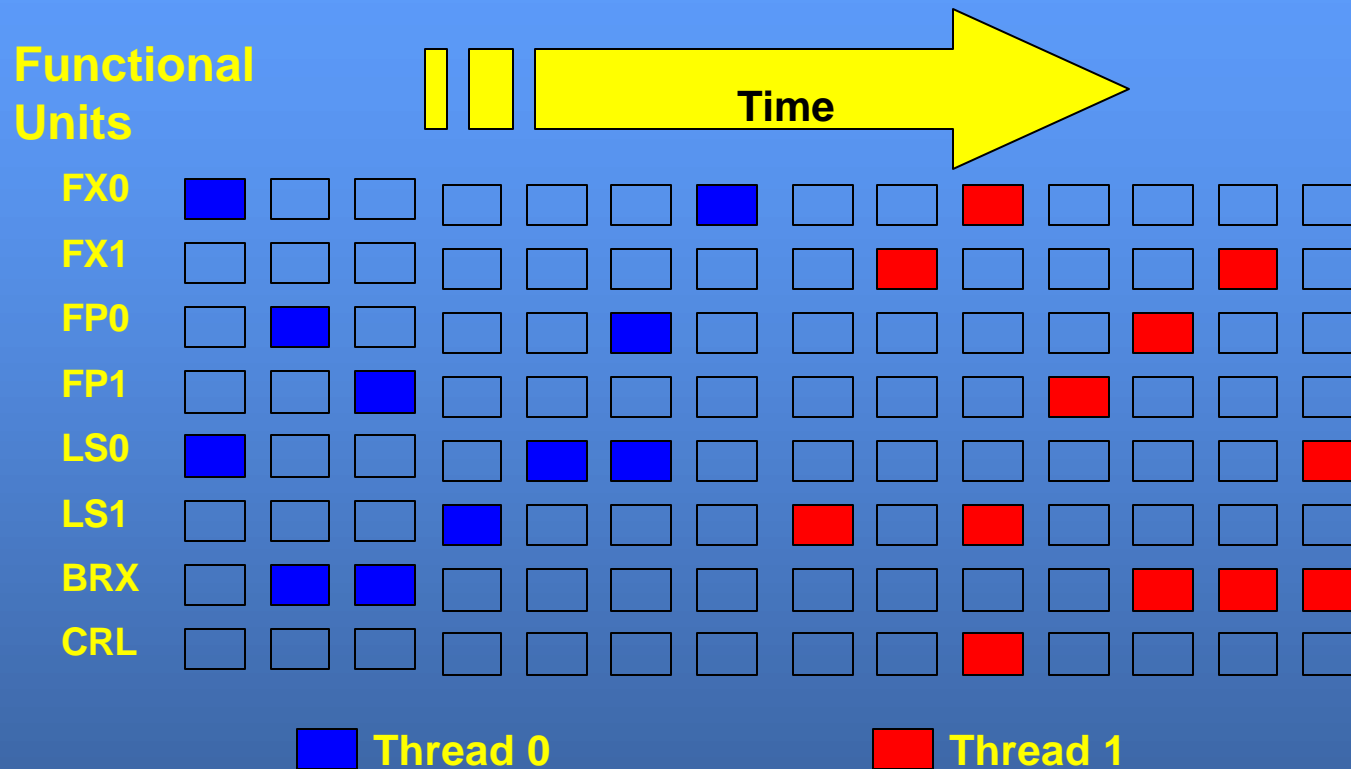


Simultaneous Multi-Threading

- **Two threads per processor**
 - Threads execute concurrently
- **Threads share:**
 - Caches
 - Registers
 - Functional units



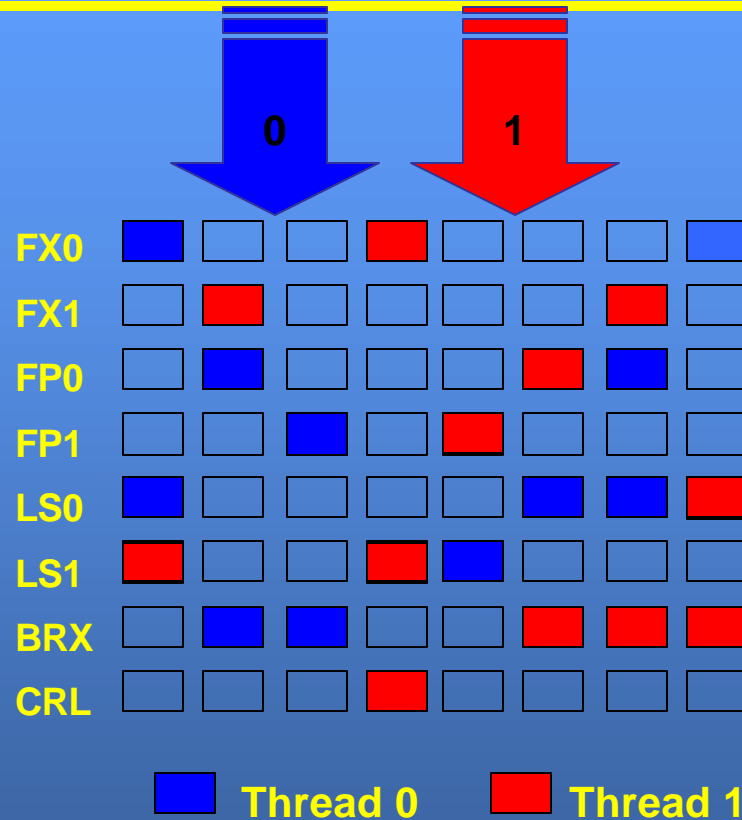
Conventional Multi-Threading



- **Threads alternate**
 - **Nothing shared**



Simultaneous Multi-Threading



- **Simultaneous execution**
 - Shared registers
 - Shared functional units



Multithreading Usage

- Program features benefiting from multithreading:
 - Functional unit latency
 - Dependencies
 - Interrupts
- Program features NOT benefiting from multithreading:
 - Bandwidth limited
 - Cache size limited

Simultaneous Multi-Threading is both a capability and a capacity feature



POWER5 Memory Bandwidth

- **Memory stores:**
 - 3x versus POWER4
- **Memory loads:**
 - 1.5x versus POWER4
- **Bandwidth dependent on:**
 - Memory configuration
 - Parts
 - DDR1
 - DDR2

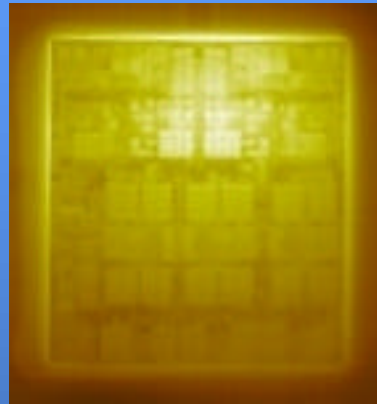


Dynamic Power Management

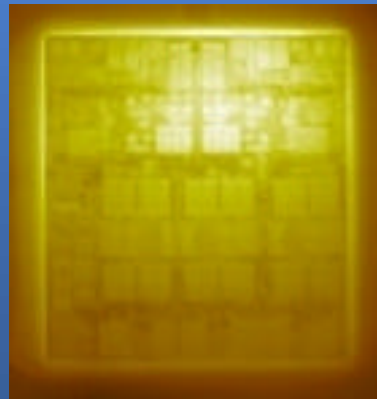
No Power Management

Dynamic Power Management

Single
Thread



Simultaneous
Multi-threading



Photos taken
with thermal
sensitive camera
while prototype
POWER5 chip
was undergoing
tests



POWER5 Summary

- **POWER4 compatibility**
- **Improved cache structure**
 - L3 on module
- **Simultaneous Multi-Threading**
- **Dynamic power management**



High Performance Switch (HPS)

- Also Known As “Federation”
- Follow on to SP Switch2
 - Also known as “Colony”
- Specifications:
 - 2 Gbyte/s (bidirectional)
 - 5 microsecond latency
- Configuration:
 - Up to four adaptors per node
 - 2 links per adaptor
 - 16 Gbyte/s per node



HPS Specifications

	Latency [microsec.]	Bandwidth, single [Mbyte/s]	Bandwidth, multiple [Mbyte/s]
Current	10	1350	1500
Expected	→ 8 → less	1400	2000



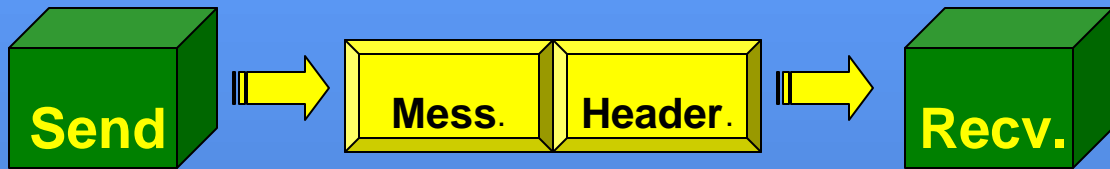
HPS Software

- **MPI-LAPI (PE V4.1)**
 - Uses LAPI as the reliable transport
 - Library uses threads, not signals for async activities
- **Existing applications binary compatible**
- **New performance characteristics**
- **New environment variables**
 - Some old ones ignored



MPI Transfer Protocols

Small Messages:
Eager

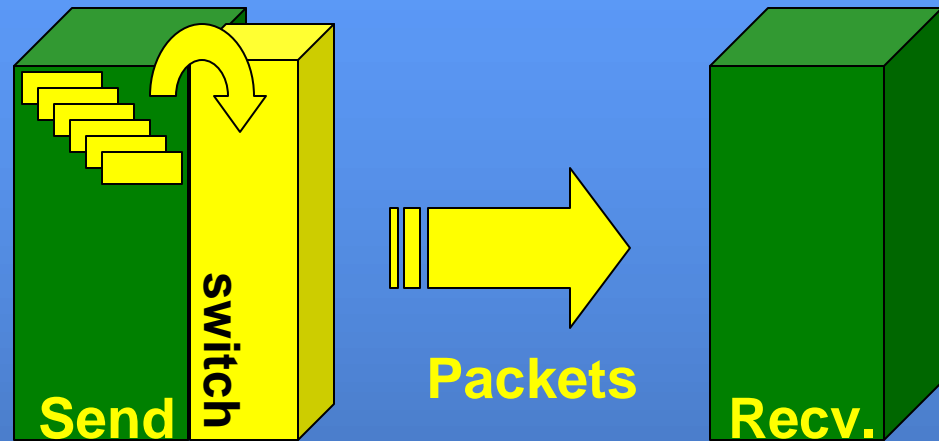


Large Messages:
Rendezvous

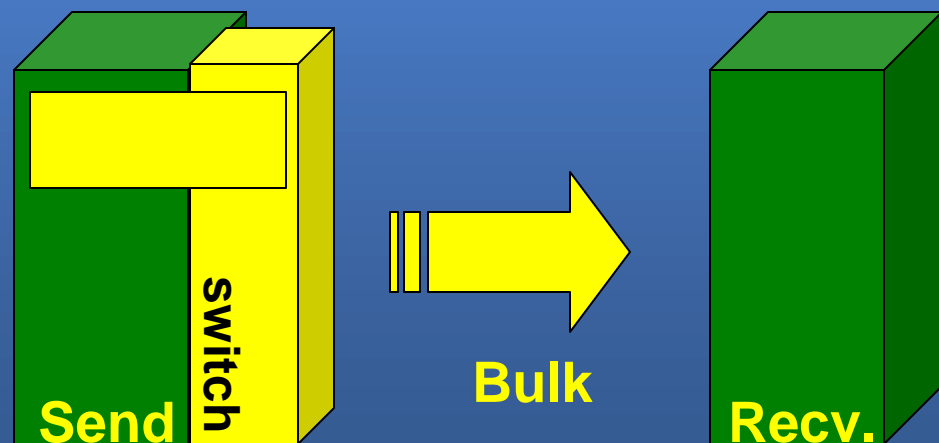


MPI Transfer Mechanisms

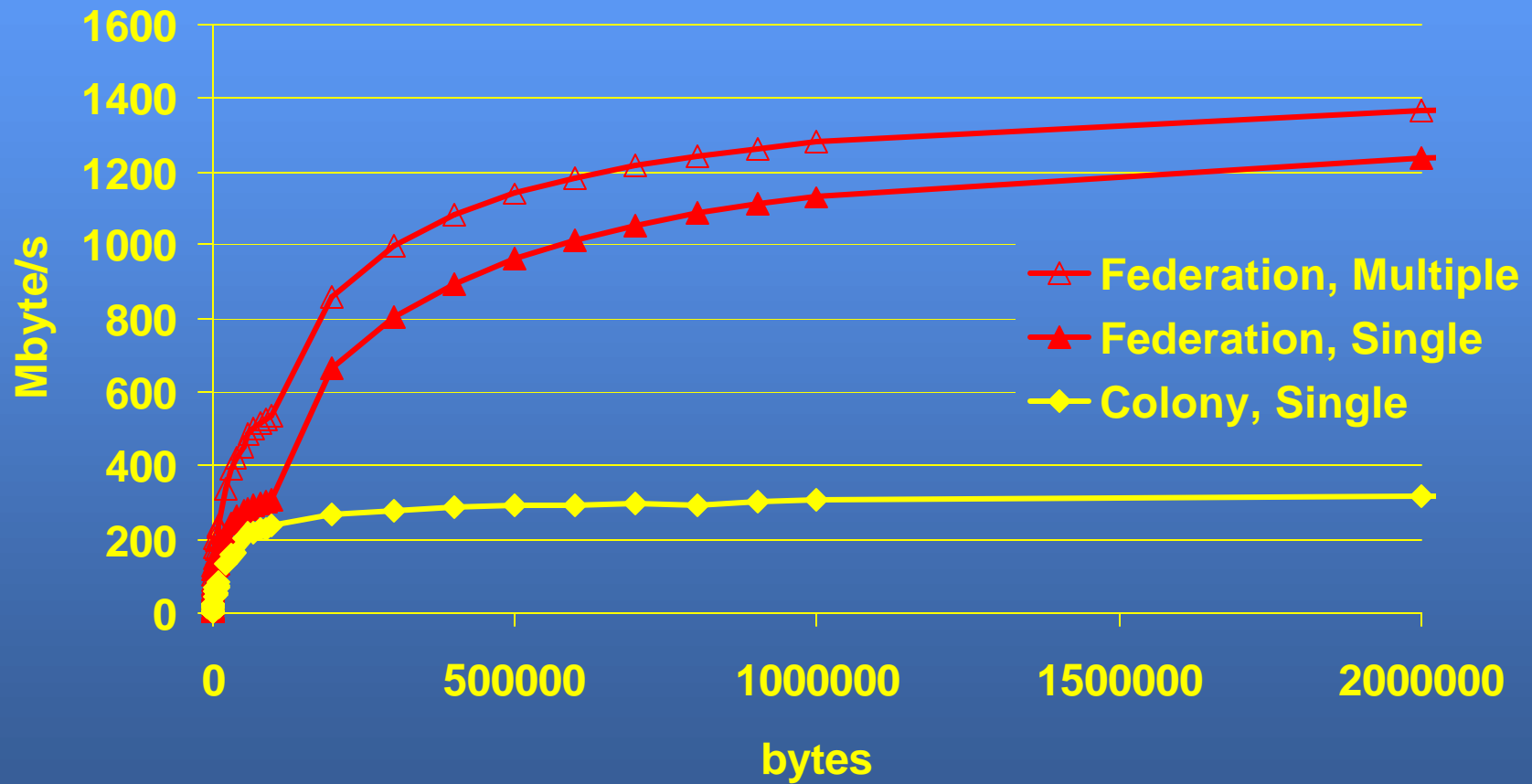
Small Messages:
Packets



Large Messages:
Bulk

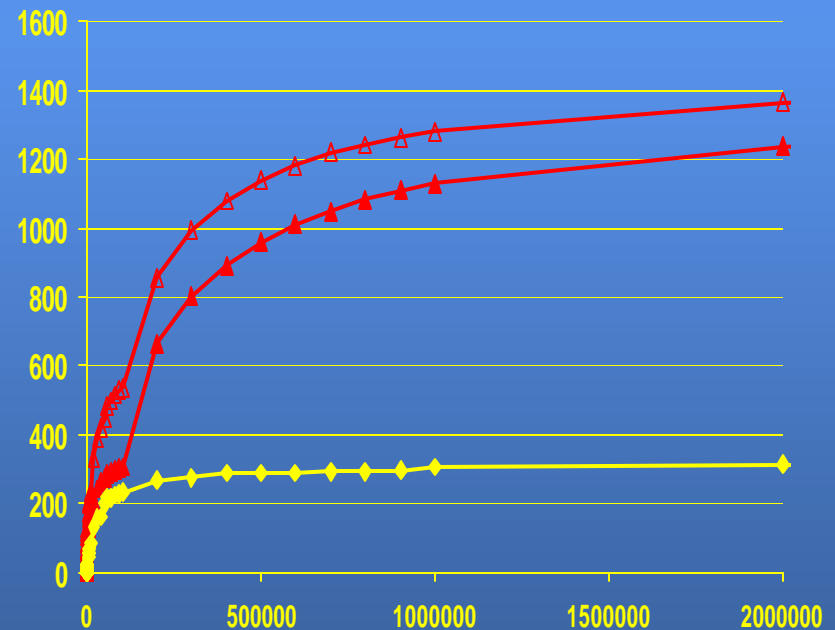


Performance: Bandwidth



HPS Performance

- High asymptotic peak bandwidth
 - ~4x vs. Colony
- Extra “kink” in performance curve
 - Bulk Transfer
- Small message performance will improve...
 - New microcode
 - Memory bandwidth limits



Bandwidth Structure: Performance Aspects

- Shared memory
- Large pages
- Bulk Transfer
- Eager Limit
- Single threaded



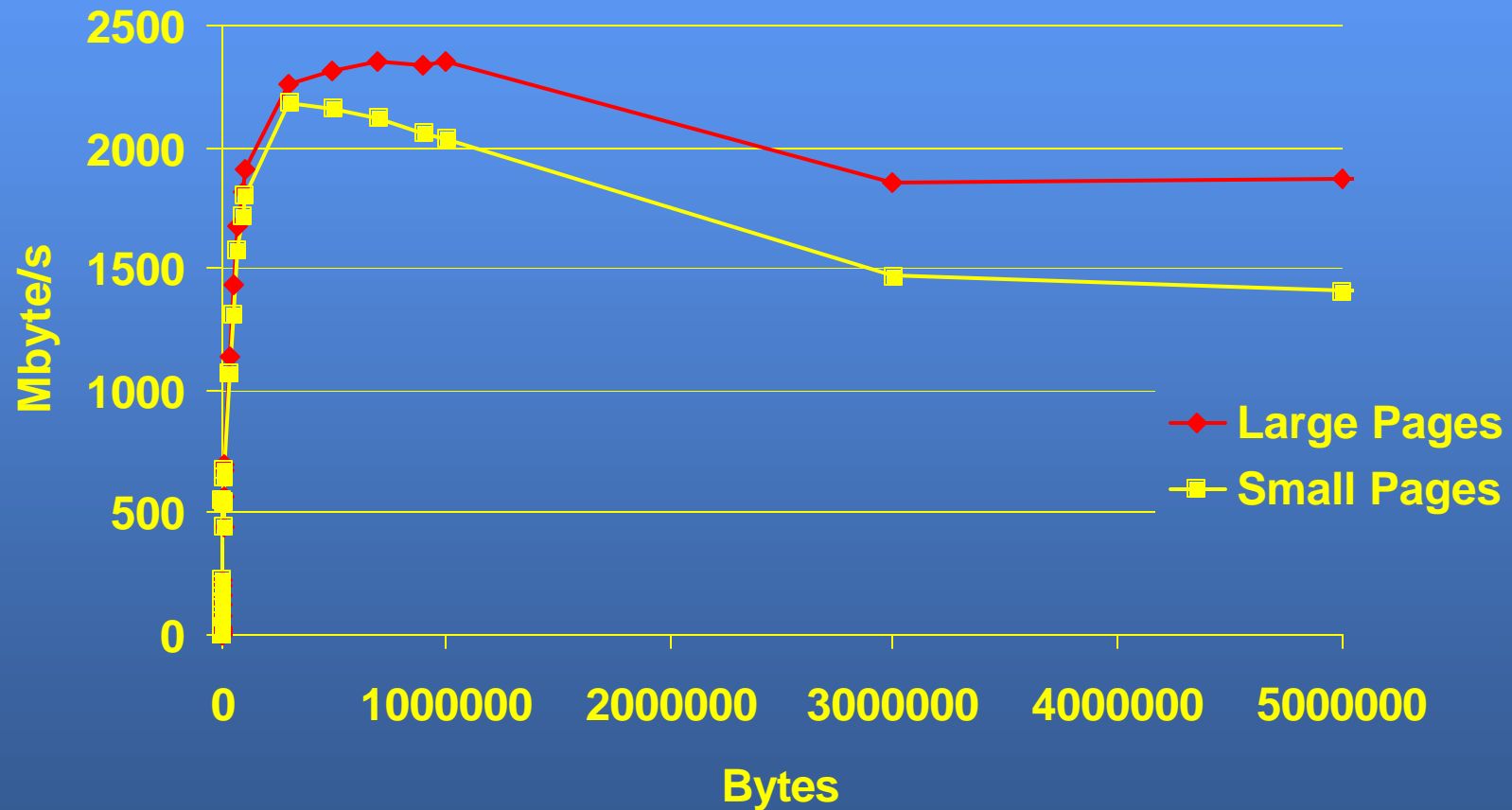
MPI Environment Variables

<i>Environment Variable</i>	<i>Recommend Value</i>
MP_EUILIB	us
MP_EUDEVICE	css0
MP_SHARED_MEMORY	yes
MP_SINGLE_THREAD	Yes*
MP_USE_BULK_XFER	yes
MP_BULK_MIN_MSG_SIZE	128000
LAPI_DEBUG_BULK_XFER_SIZE	1000000
LDR_CNTRL	LARGE_PAGE_DATA=Y

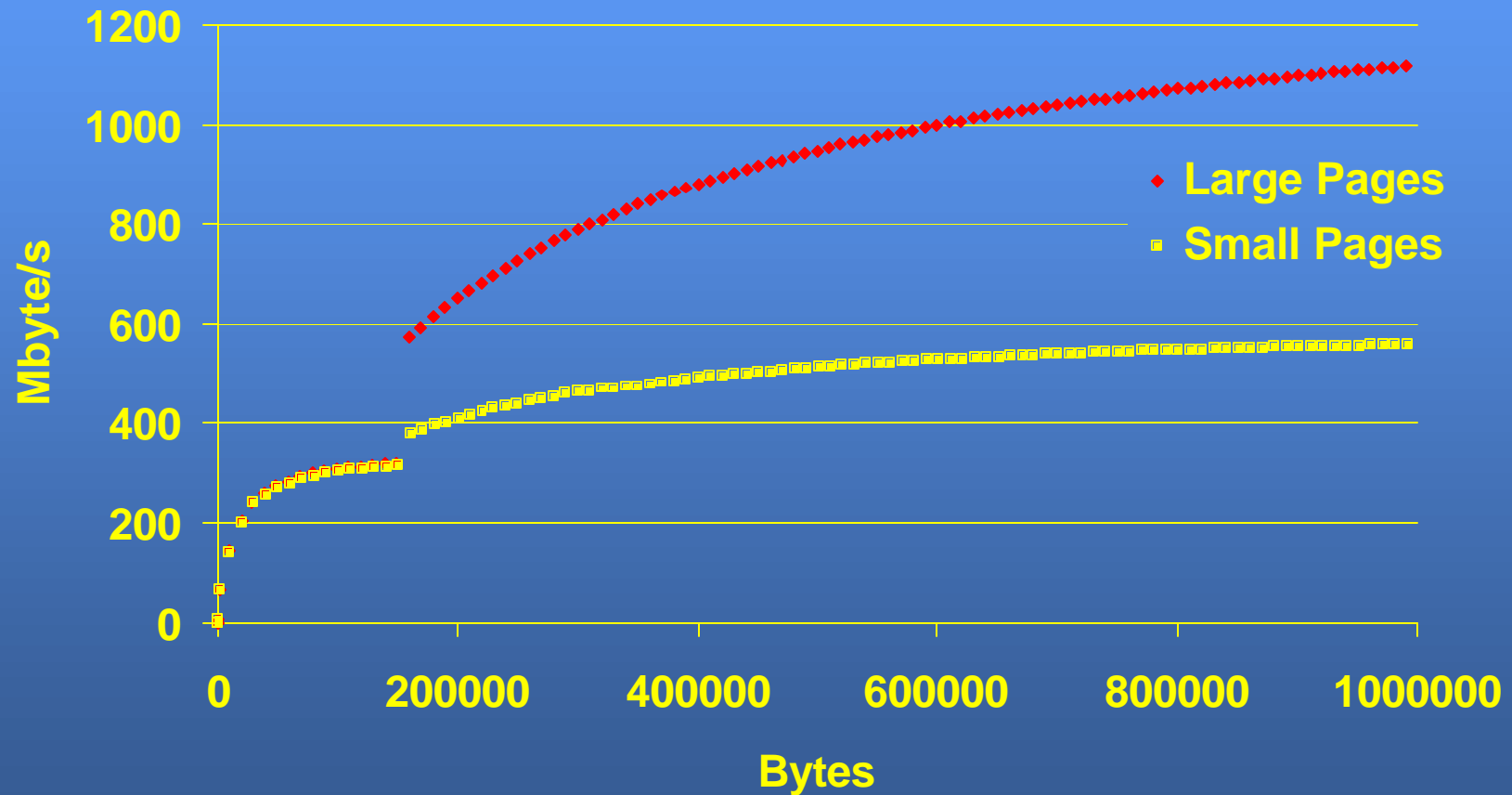
* If possible



Large Pages: Single Node



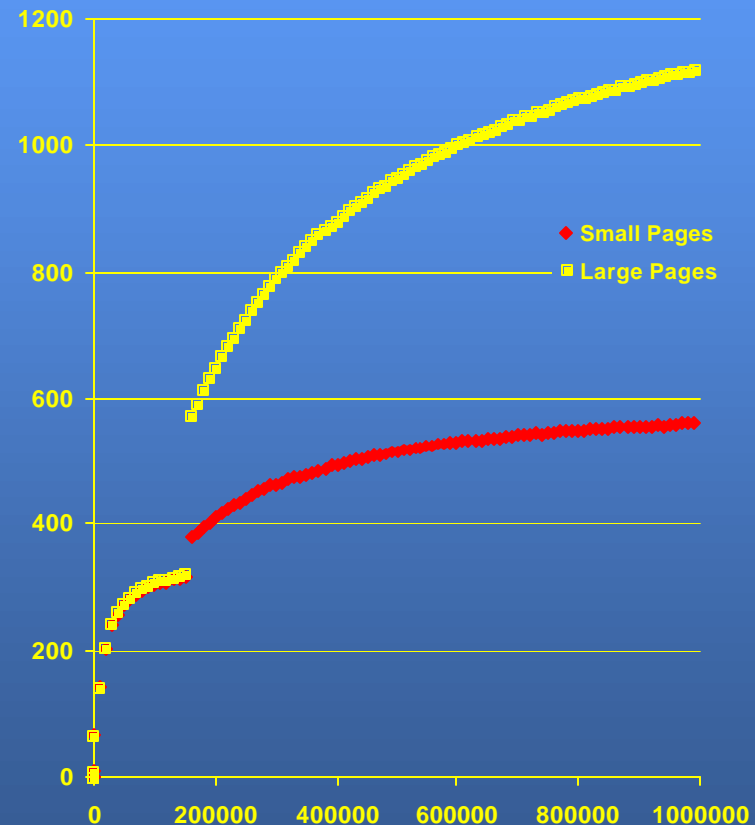
Large Pages: Inter-node



IBM	Business Partner	IBM	Business Partner
IBM	Business Partner	IBM	Business Partner
IBM	Business Partner	IBM	Business Partner
IBM	Business Partner	IBM	Business Partner
IBM	Business Partner	IBM	Business Partner

Large Pages

- **Controlled by application**
 - -blpdata
 - LDR_CNTL
LARGE_PAGE_DATA=
{MNY}
- **Significant performance increase**
 - ~ 2x



Bulk Transfer

- High bandwidth mechanism
- Latency is ~150 microseconds
 - Default start is 128 kbyte
 - MP_BULK_MIN_MSG_SIZE
- Default transfer size: 1 Mbyte
 - LAPI_DEBUG_BULK_XFER_SIZE



Bulk Transfer

- Bandwidth, and “latency”, change at size 128 kbyte (default)
- Environment variable:
 - MP_BULK_MIN_MSG_SIZE

