

HPS Usage and Performance

ScicomP9
March, 2004

Charles Grassl
IBM

1
© 2003 IBM Corporation



Agenda

- High Performance Switch

2
© 2003 IBM Corporation



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

3
© 2003 IBM Corporation



HPS Specifications

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

4
© 2003 IBM Corporation



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

5
© 2003 IBM Corporation

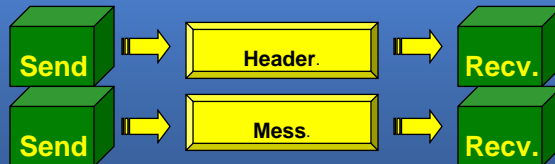


MPI Transfer Protocols

Small Messages:
Eager



Large Messages:
Rendezvous

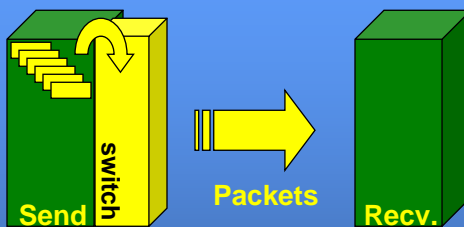


6
© 2003 IBM Corporation

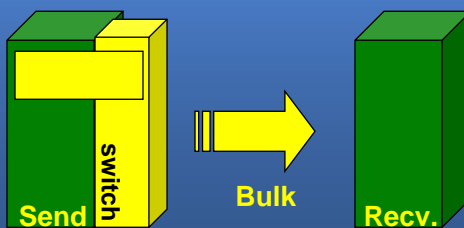


MPI Transfer Mechanisms

Small Messages:
Packets

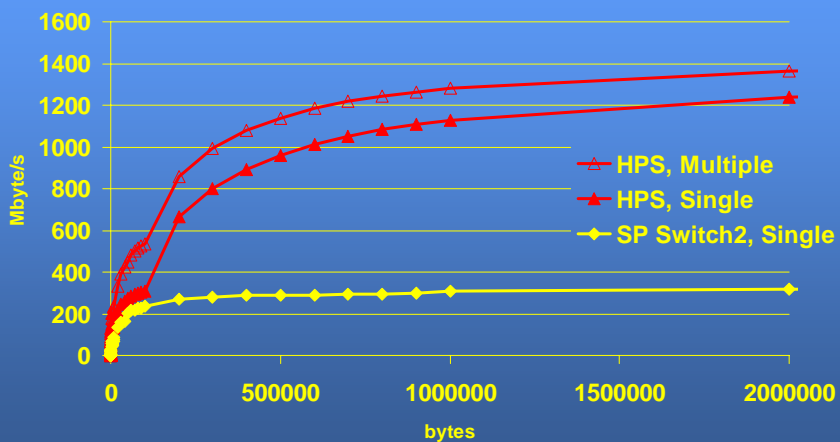


Large Messages:
Bulk



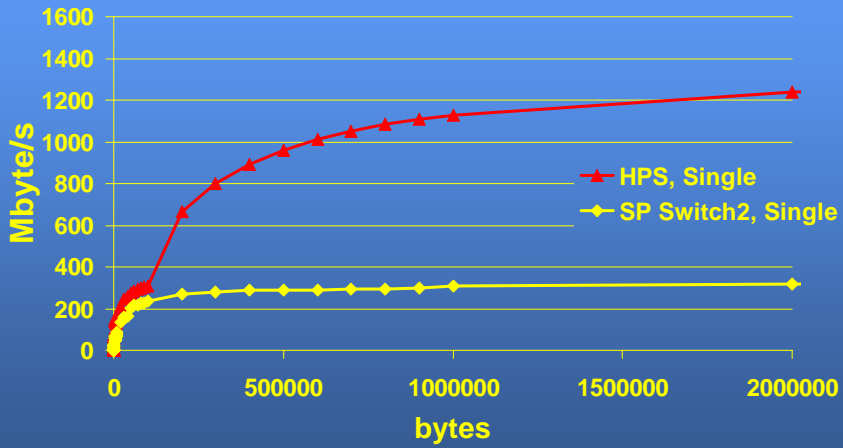
7
© 2003 IBM Corporation

Performance: Bandwidth



8
© 2003 IBM Corporation

Performance: Bandwidth

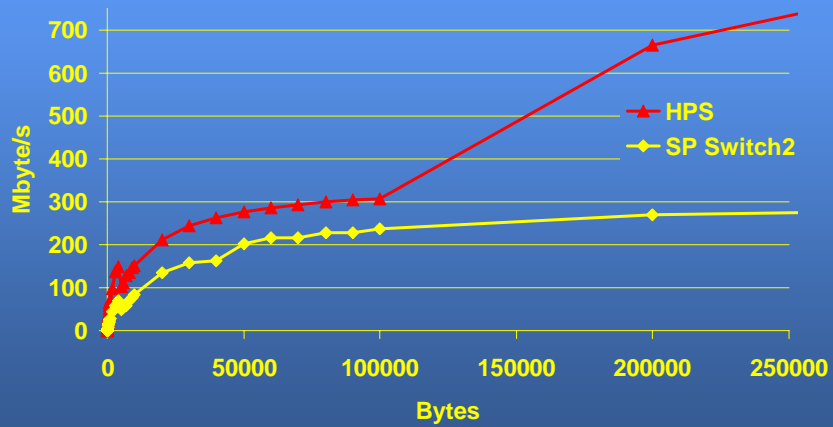


9

© 2003 IBM Corporation



Performance: Bandwidth

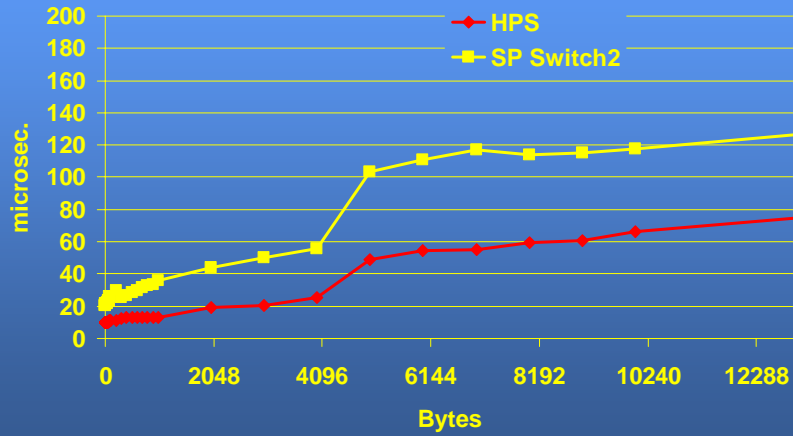


10

© 2003 IBM Corporation



Performance: Latency Eager Limit Crossover

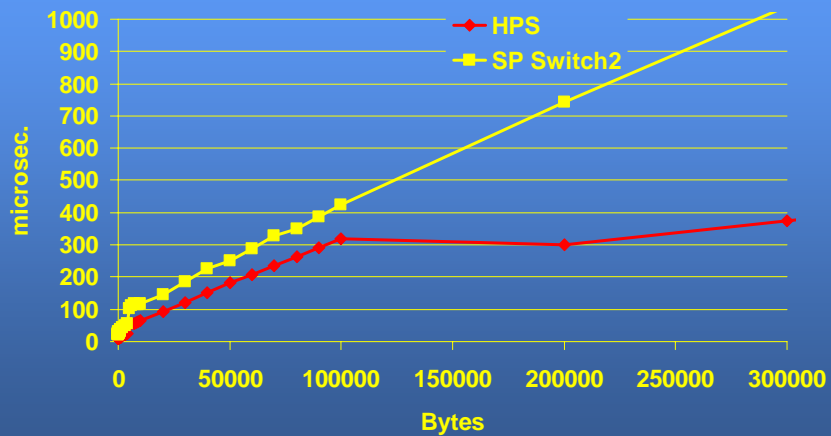


11

© 2003 IBM Corporation



Performance: Latency Bulk Transfer Crossover



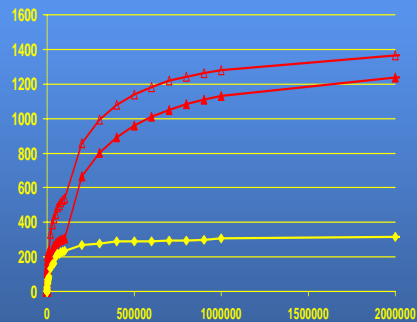
12

© 2003 IBM Corporation



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



13

© 2003 IBM Corporation



Bandwidth Structure: Performance Aspects

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

14

© 2003 IBM Corporation



MPI Environment Variables

<i>Environment Variable</i>	<i>Recommend Value</i>
MP_EUILIB	us
MP_EUIDEVICE	css0, sn_single
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

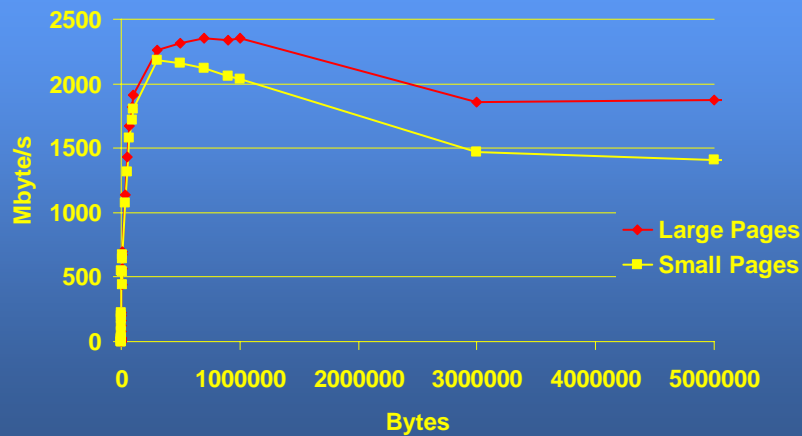
* If possible

15

© 2003 IBM Corporation



Large Pages: Single Node

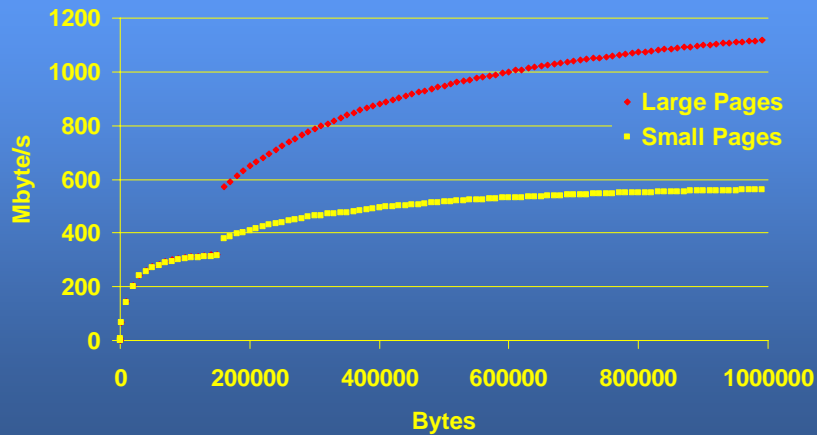


16

© 2003 IBM Corporation



Large Pages: Inter-node



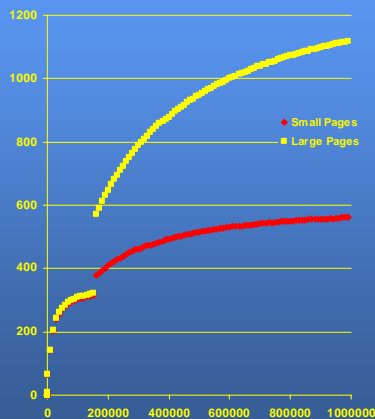
17

© 2003 IBM Corporation



Large Pages

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



18

© 2003 IBM Corporation



Bulk Transfer

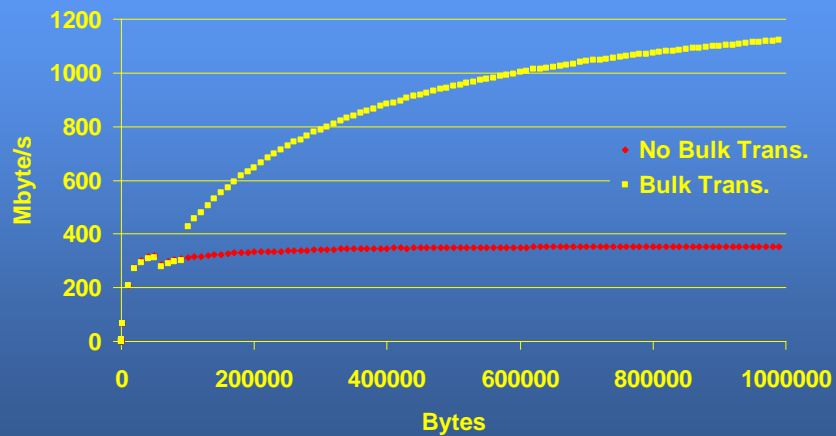
- High bandwidth mechanism
- Latency is ~150 microseconds
 - Default start is 128 kbyte
 - MP_BULK_MIN_MSG_SIZE
- Default transfer size: 1 Mbyte
 - LAPI_DEBUG_BULK_XFER_SIZE

19

© 2003 IBM Corporation



Bulk Transfer



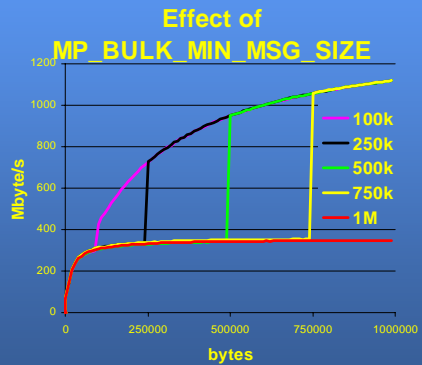
20

© 2003 IBM Corporation



Bulk Transfer

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

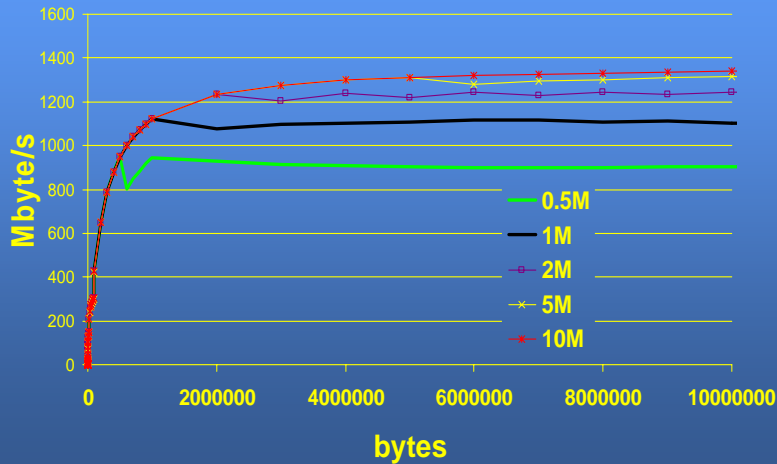


21

© 2003 IBM Corporation



Bulk Transfer Size: LAPI_DEBUG_BULK_XFER_SIZE



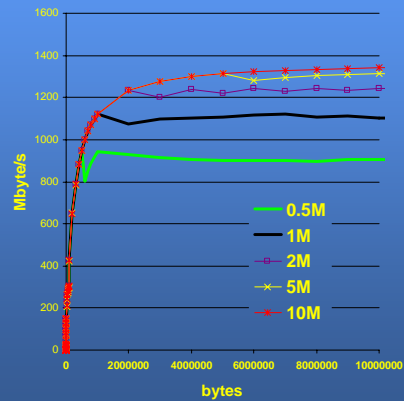
22

© 2003 IBM Corporation



Bulk Transfer Size

- Affects bandwidth for very large messages
- Environment variable:
 - LAPI_DEBUG_BULK_XFER_SIZE
 - Default: 1 M



23

© 2003 IBM Corporation



Summary

- HPS
 - Bandwidth
 - Bulk Transfers
 - 4x higher bandwidth (large messages)
 - Latency
 - $\frac{1}{2}$ latency \rightarrow $\frac{1}{4}$

24

© 2003 IBM Corporation

