HPS Usage and Performance

ScicomP9
March, 2004

Charles Grassl
IBM

Agenda

- **High Performance Switch**

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

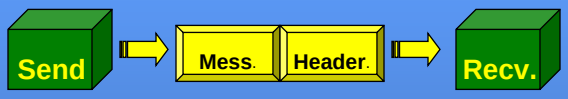
	Latency [microsec.]	Bandwidth, single [Mbyte/s]	Bandwidth, multiple [Mbyte/s]
Current	10	1350	1500
Expected	→ 8 → less	1400	2000

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

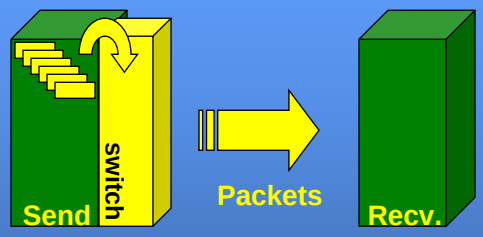
Small Messages: Eager



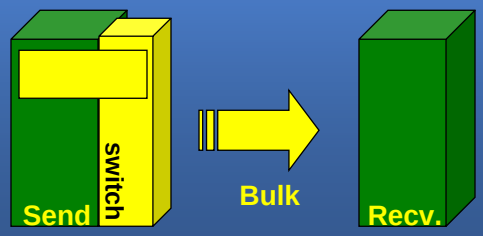
Large Messages: Rendezvous

MPI Transfer Mechanisms

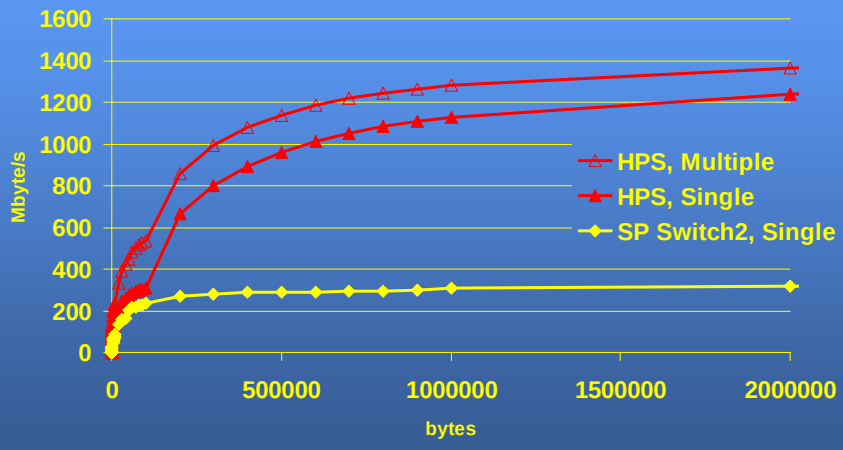
Small Messages:
Packets



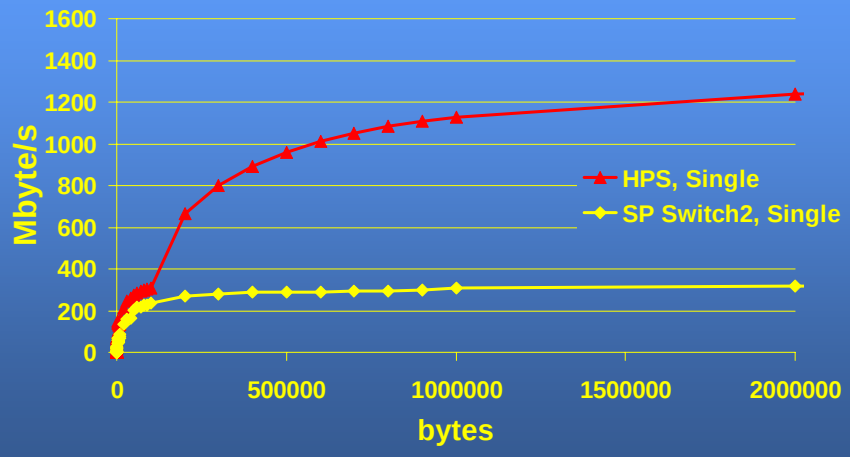
Large Messages:
Bulk



Performance: Bandwidth



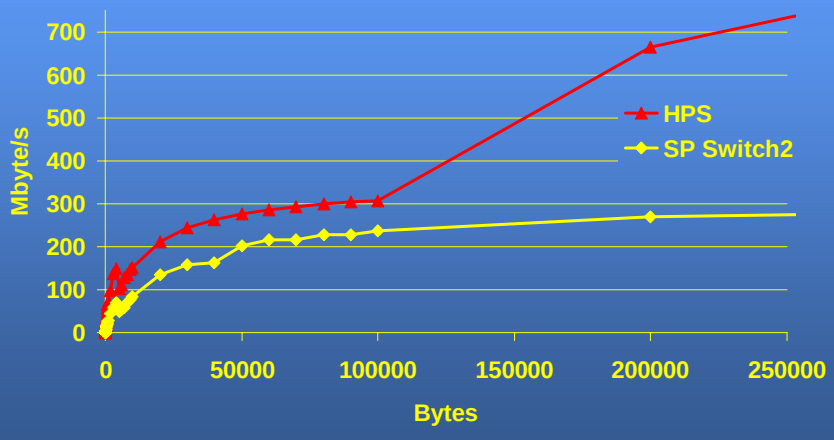
Performance: Bandwidth



9

IBM

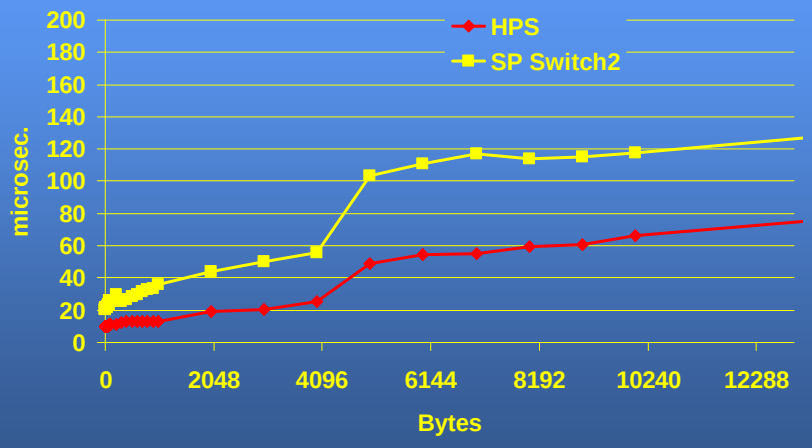
Performance: Bandwidth



10

IBM

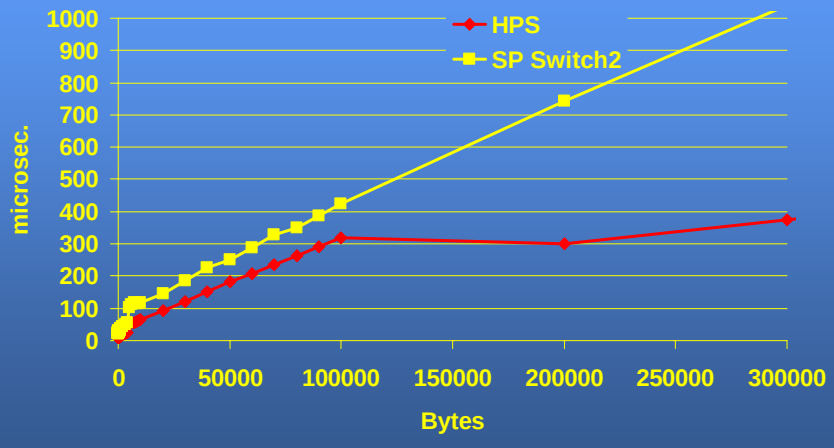
Performance: Latency Eager Limit Crossover



11

IBM

Performance: Latency Bulk Transfer Crossover

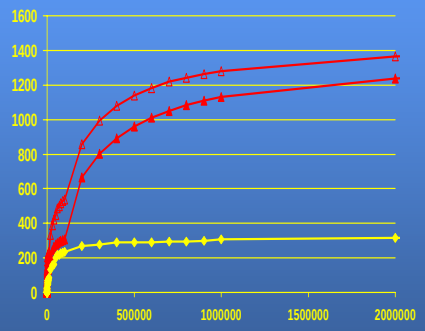


12

IBM

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

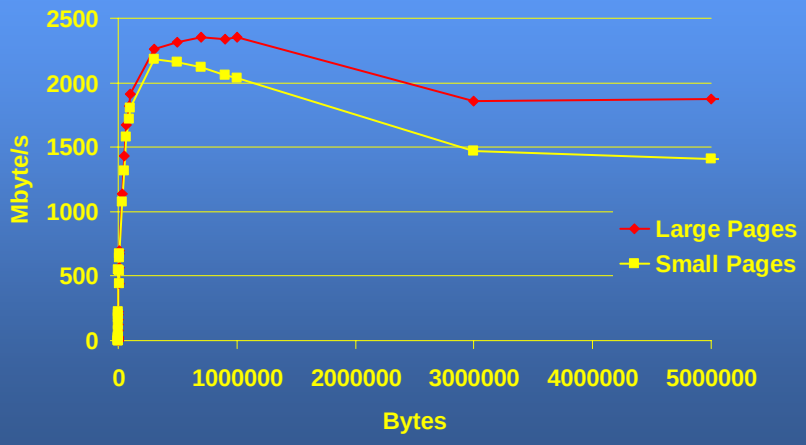
Environment Variable	*Recommend Value*
MP_EUILIB	us
MP_EUIDEVICE	css0, sn_single
MP_SHARED_MEMORY	yes
MP_SINGLE_THREAD	Yes*
MP_USE_BULK_XFER	yes
MP_BULK_MIN_MSG_SIZE	128000
LAPI_DEBUG_BULK_ XFER_SIZE	1000000
LDR_CNTRL	LARGE_PAGE_DATA=Y

15

* If possible

IBM

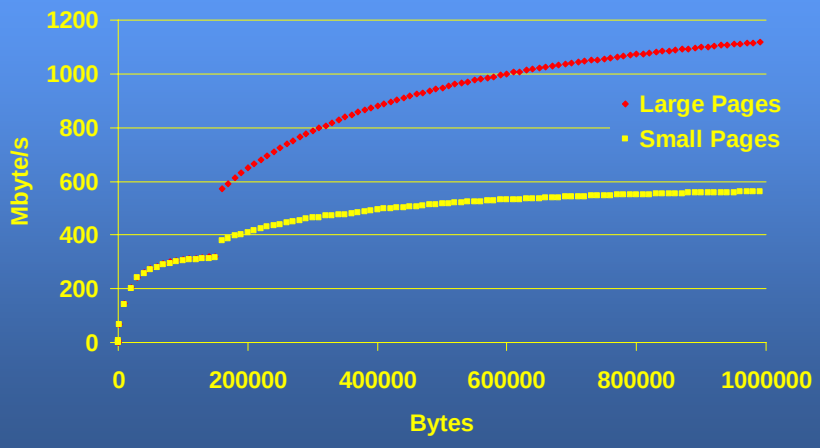
Large Pages: Single Node



16

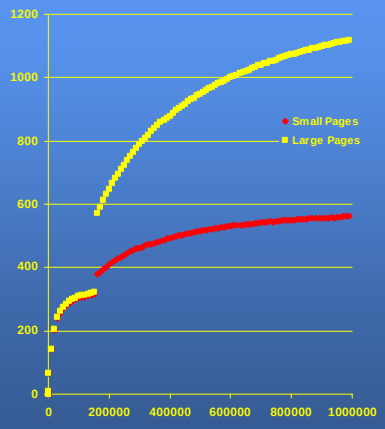
IBM

Large Pages: Inter-node



Large Pages

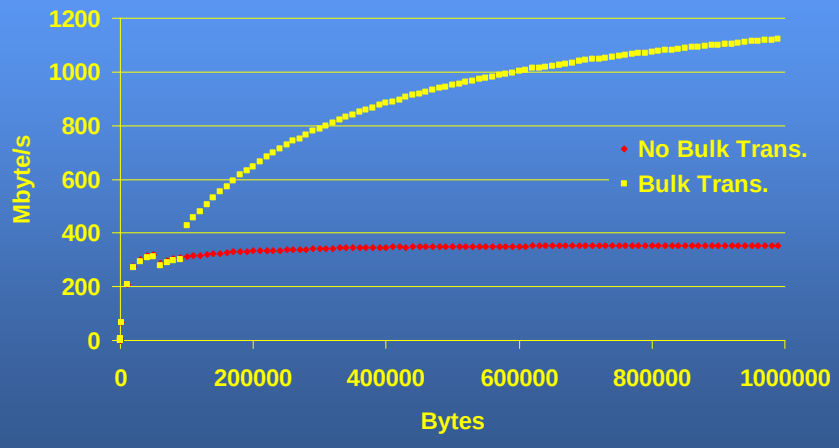
- Controlled by application
  - -blpdata
  - LDR_CNTL LARGE_PAGE_DATA= {MNY}
- Significant performance increase
  - ~ 2x



Bulk Transfer

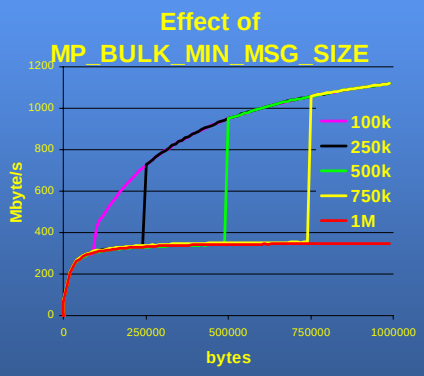
- High bandwidth mechanism
- Latency is ~150 microseconds
  - Default start is 128 kbyte
  - MP_BULK_MIN_MSG_SIZE
- Default transfer size: 1 Mbyte
  - LAPI_DEBUG_BULK_XFER_SIZE

Bulk Transfer

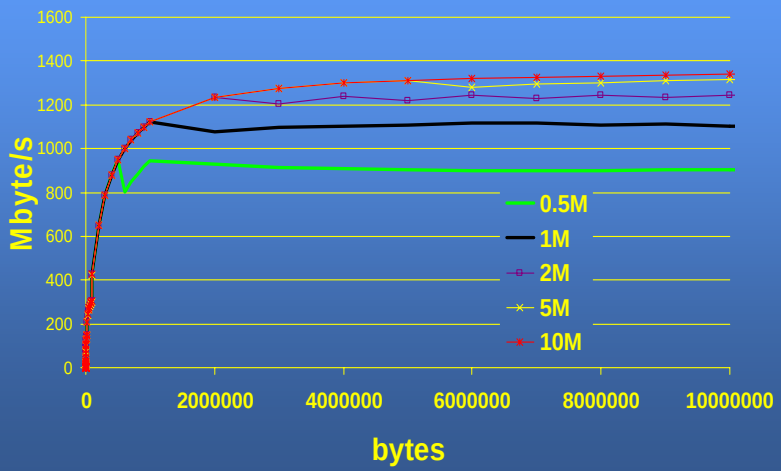


Bulk Transfer

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

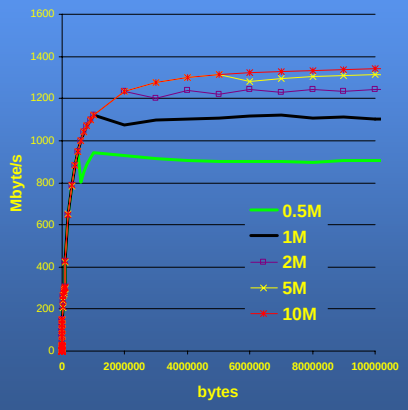


Bulk Transfer Size:
LAPI_DEBUG_BULK_XFER_SIZE



Bulk Transfer Size

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



Summary

- HPS
  - Bandwidth
    - Bulk Transfers
    - 4x higher bandwidth (large messages)
  - Latency
    - ½ latency → ¼