New Polling Strategy: Memory-Based Notifcation

- Old: polling for new packet
  - ☆ CPU->fm polls LANai mem-> PCI bus
- New: Poll memory location, LCP <-> FM handshake.
  - Advantages:
    - ☆ subsequent polls satisfied in cache
    - ☆ don't use CPU/Memory or CPU/PCI buses
    - ☆ don't interfere with DMA transfers
    - ☆ uses hardware cache coherence
    - ☆ align packet to bottom of frame: poll same location and know when DMA is over
    - ☆ 15% improvement over previous HPVM (10.4 μs vs 8.8 μs)
    - ☆ supporting multiple NICs

Shared Memory Transport

- Don’t go through NIC for Intra-node IPC
- Array based data structures
- Process controls a message queue
- fetch-and-increment atomic operation
- Ring buffer
- SMT improves intranode bandwidth and latency

Self Tuning I/O code: PCI Chipset Pacing

- Different PCI bus implementations differ in performance
- Adaptive selection of packet burst-size for optimum PIO performance
- pace packets: messages > optimal packet size are broken into chunks

➠ 160 bytes for 440BX

Conclusions

- HPVM, FM
- Performance Enhancements in FM
  - New polling scheme
  - Shared memory transport
  - Intra-node enhancements
  - Adaptive tuning: Pacing PCI chipset
- Demonstrated Supercluster n > 500
- 5TB fileserver @ 1GB/s sustained network bandwidth