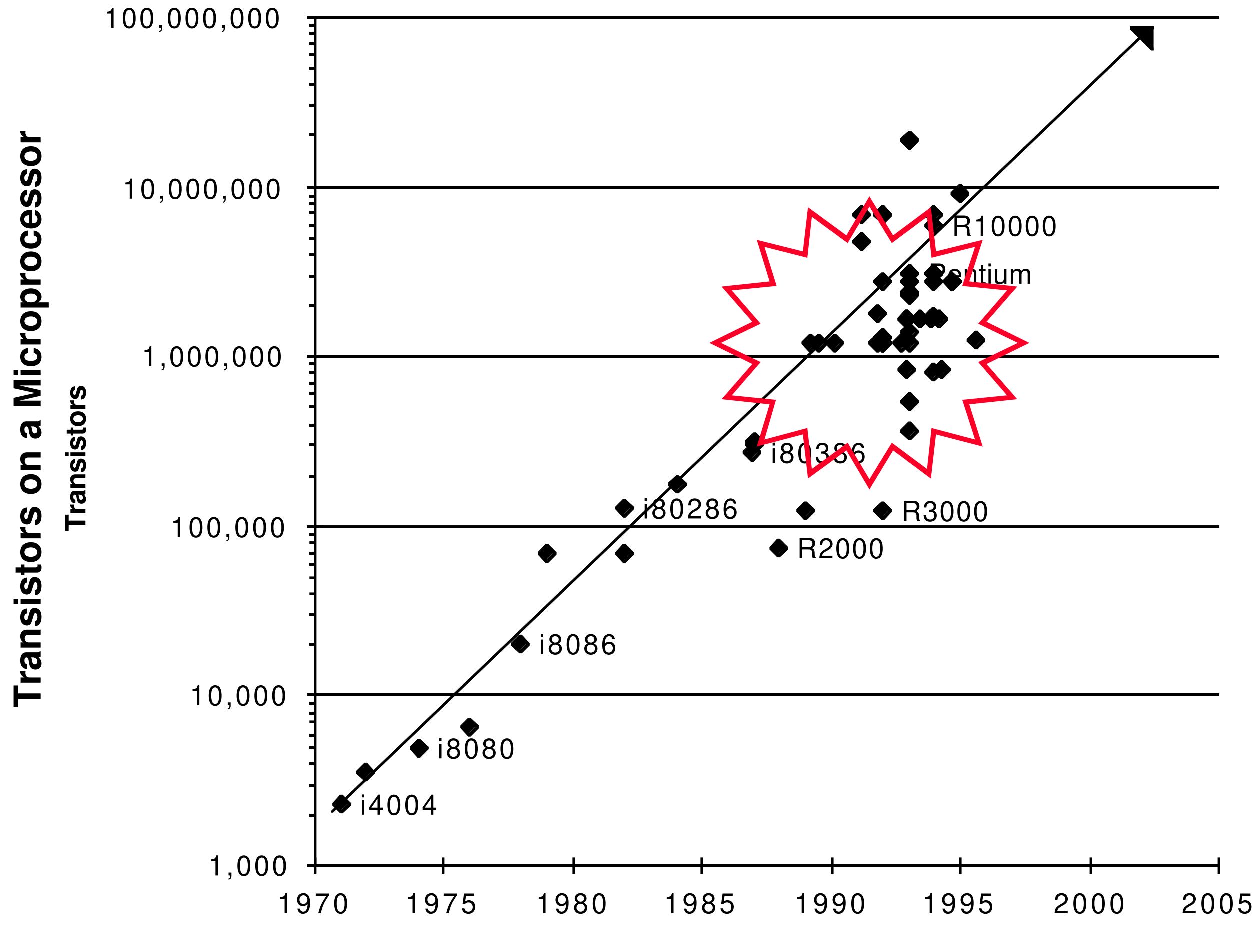
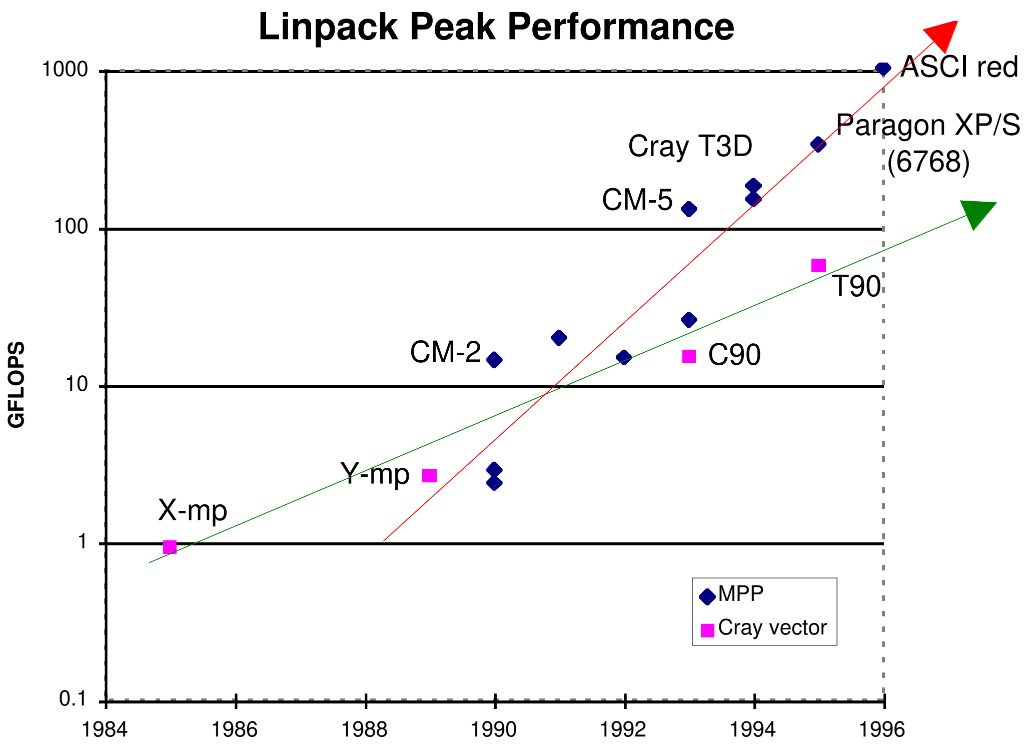
Remember the "Killer Micro" ?



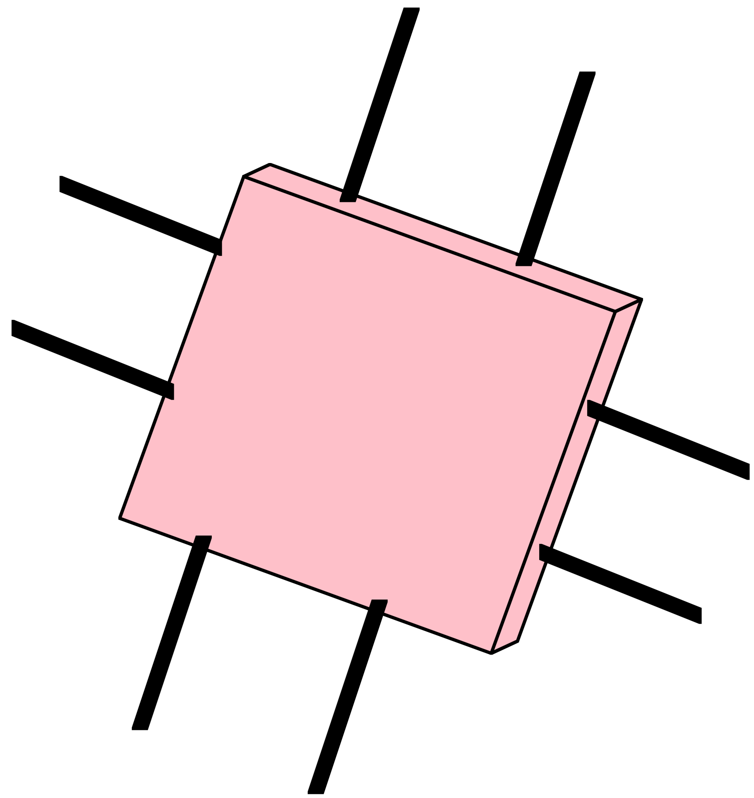
What happened at the high end?



- **Technology change in all markets**
- **At many levels: Arch, Compiler, OS, Application**

Another Tech. Revolution Brewing

- **The "Killer Switch"**
  - single chip building block for scalable networks
  - high bandwidth
  - low latency
  - very reliable
    - » if it's not unplugged



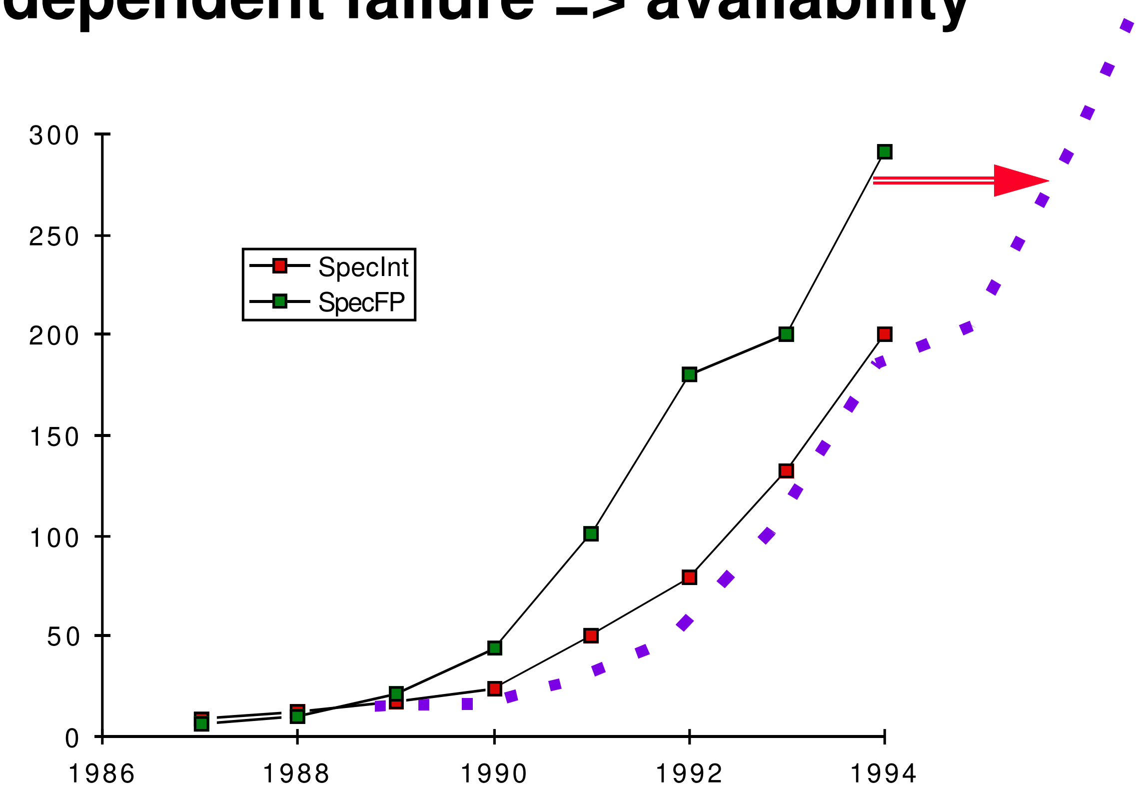
=> System Area Networks

One Example: Myrinet

- 8 bidirectional ports of 160 MB/s each way
- < 500 ns routing delay
- No smarts - just move bits
- Detects connectivity and deadlock

Potential: Snap together large systems

- incremental scalability
- time / cost to market
- independent failure => availability

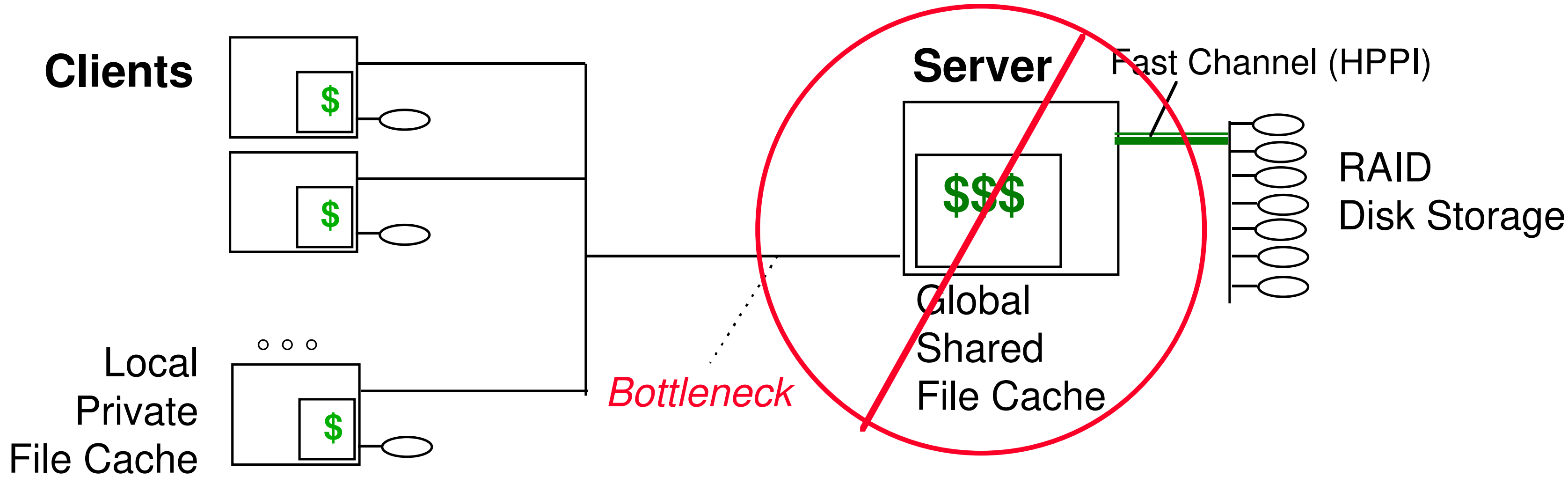


Opportunity: Rethink traditional systems design

- **Remote memory and processor are closer to you than your own disks!**

- **File system design ?**
- **Virtual Memory ?**
- **Communication Stacks ?**

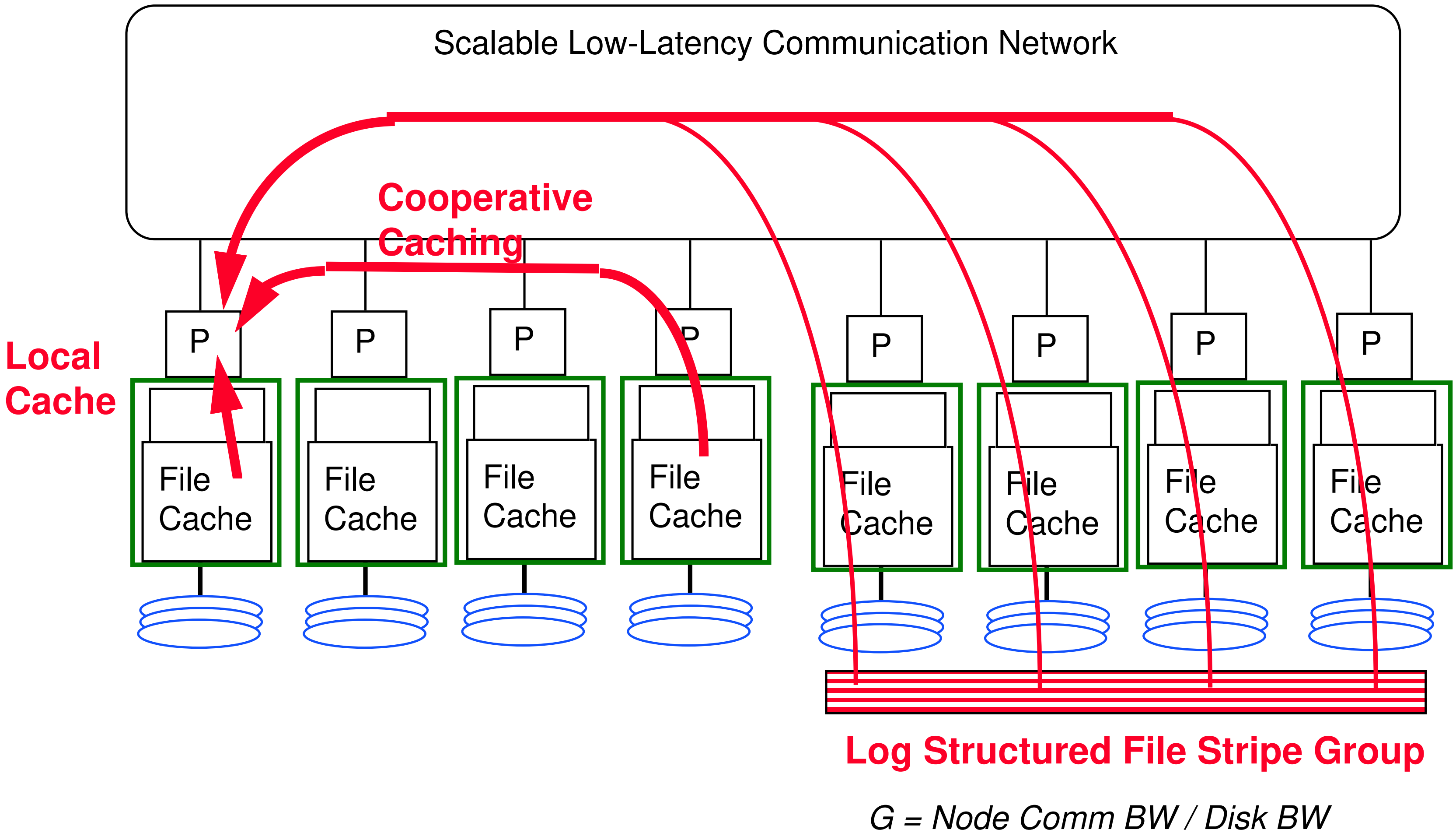
Example: Traditional File System



- **Expensive**
- **Complex**
- **Non-Scalable**
- **Single point of failure**

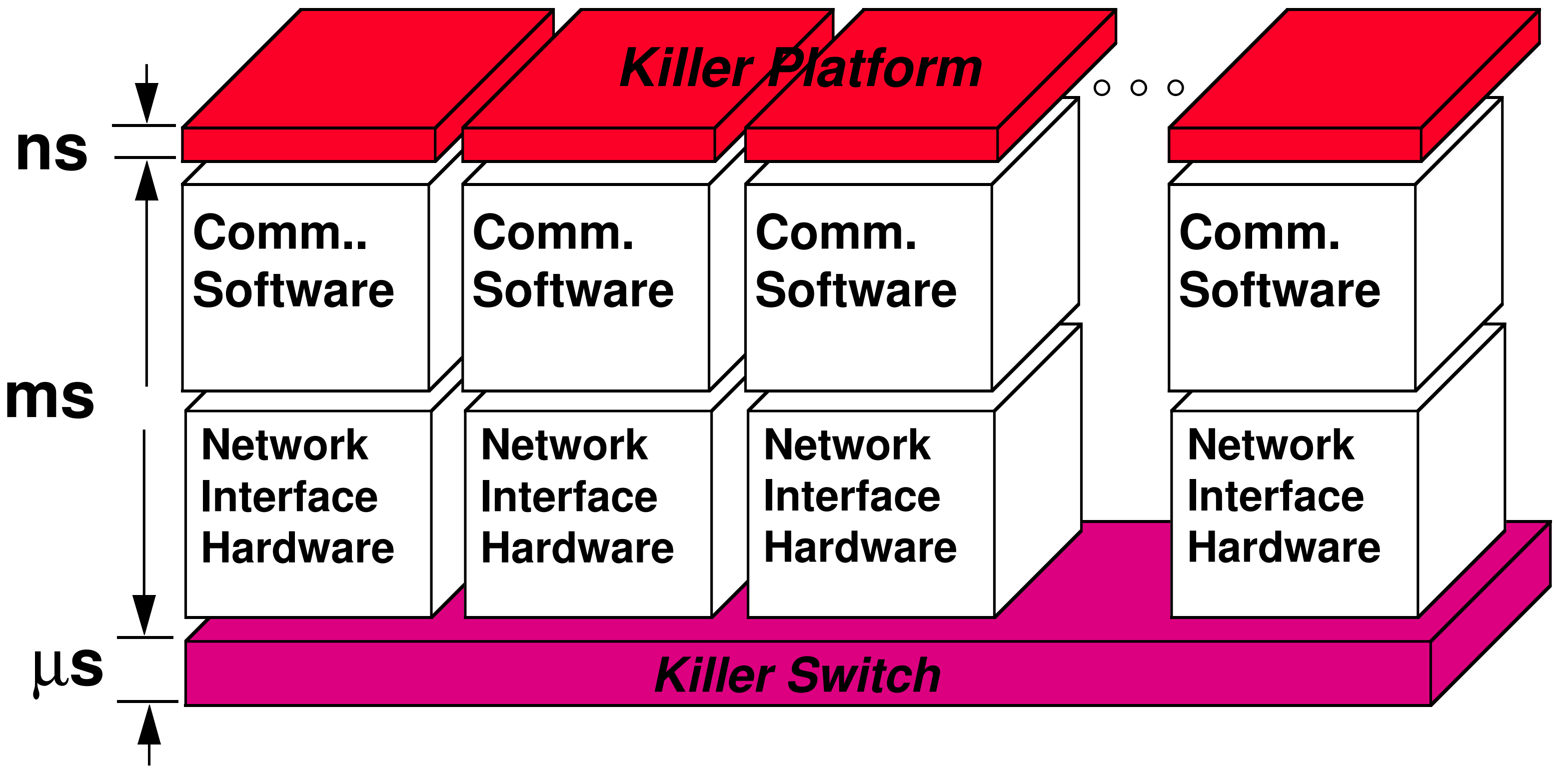
- **Server resources at a premium**
- **Client resources poorly utilized**

xFS: Truly Distributed File System



- **net VM: page to remote memory**

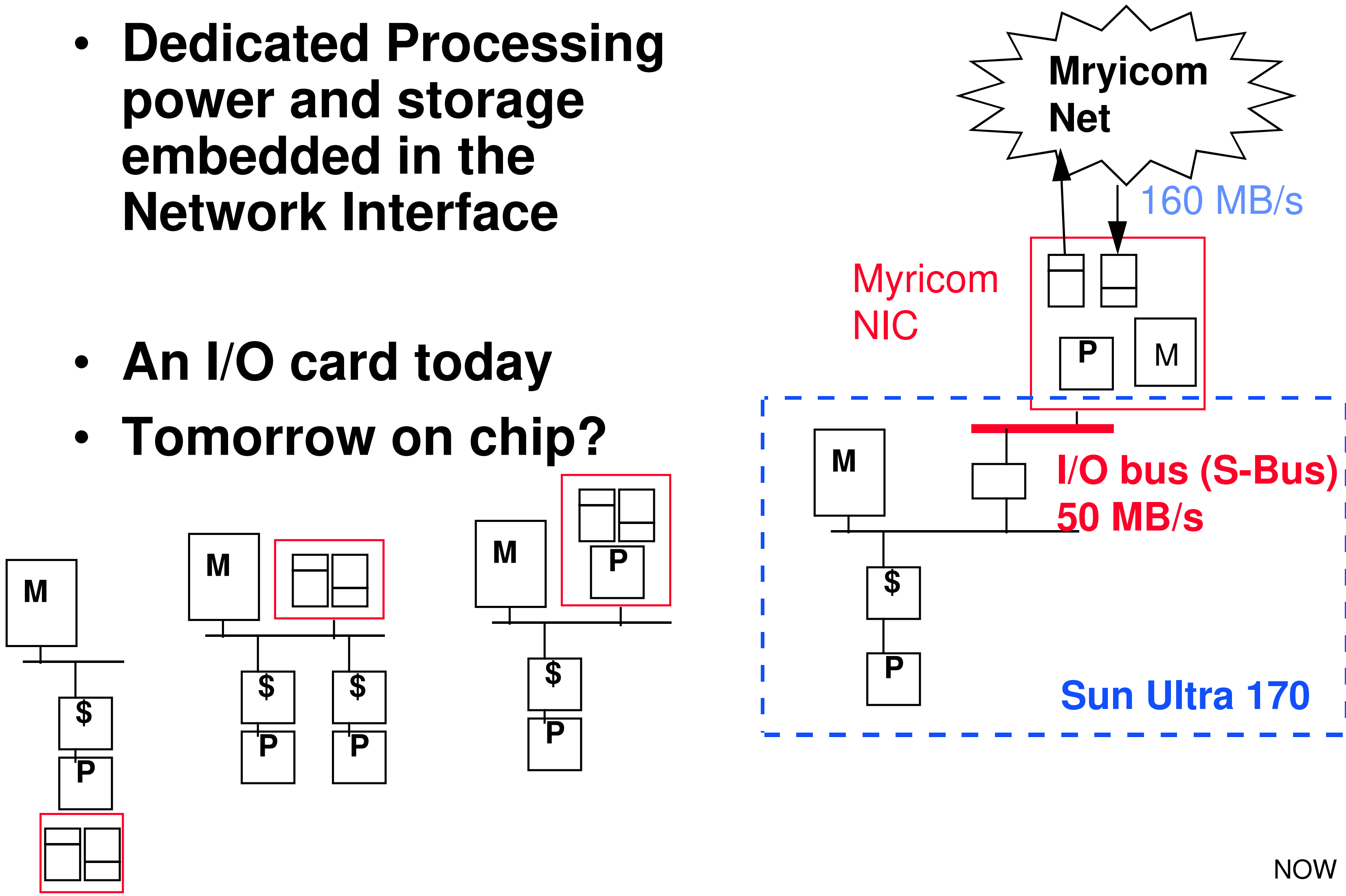
Problem: the last two inches



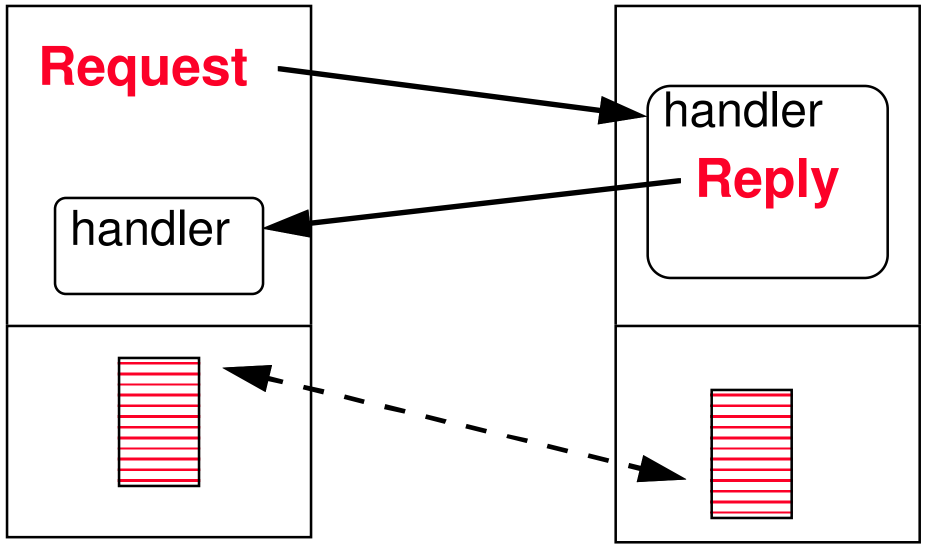
Opening: Intelligent Network Interfaces

- **Dedicated Processing power and storage embedded in the Network Interface**

- **An I/O card today**
- **Tomorrow on chip?**



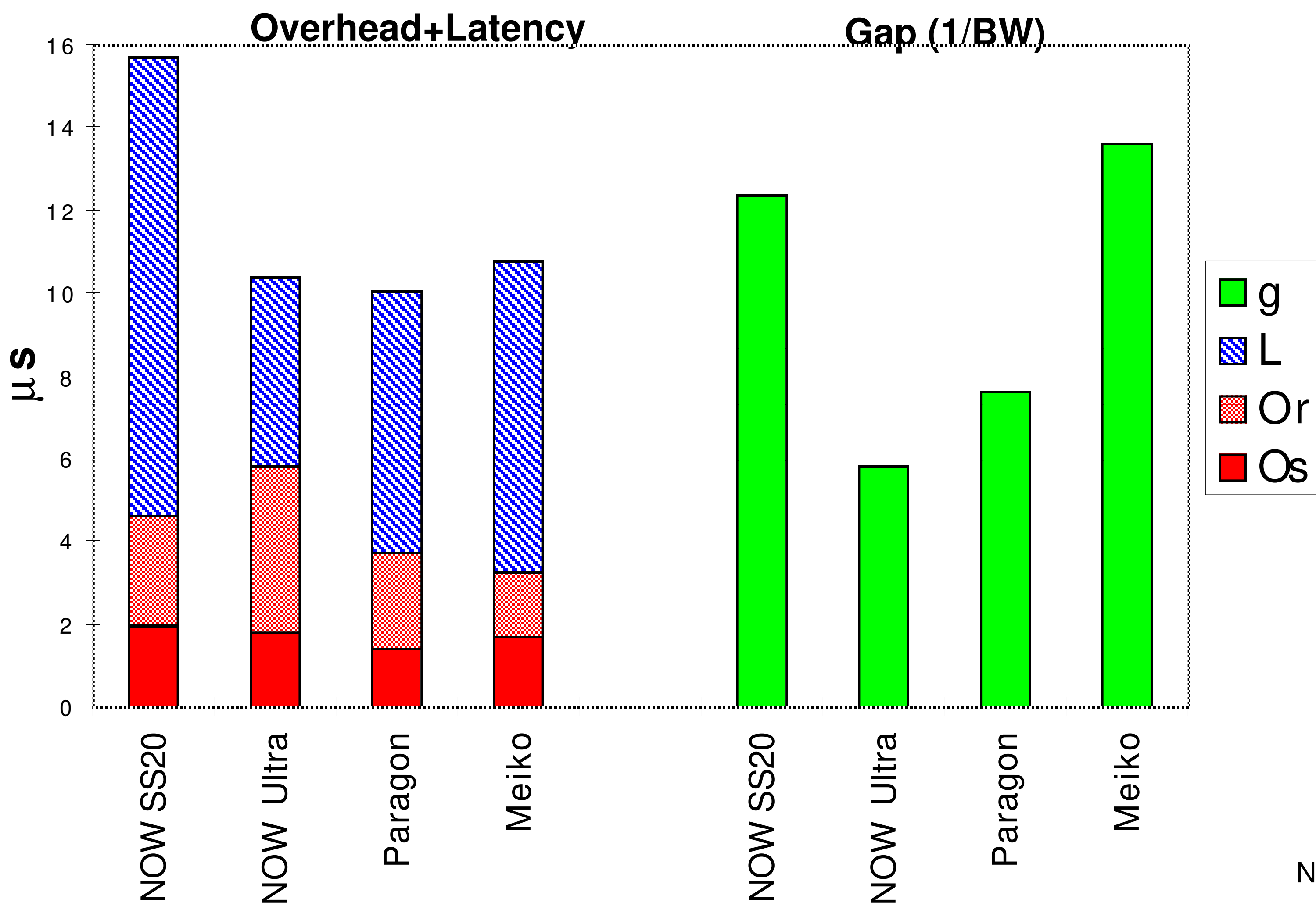
Our Attack: Active Messages



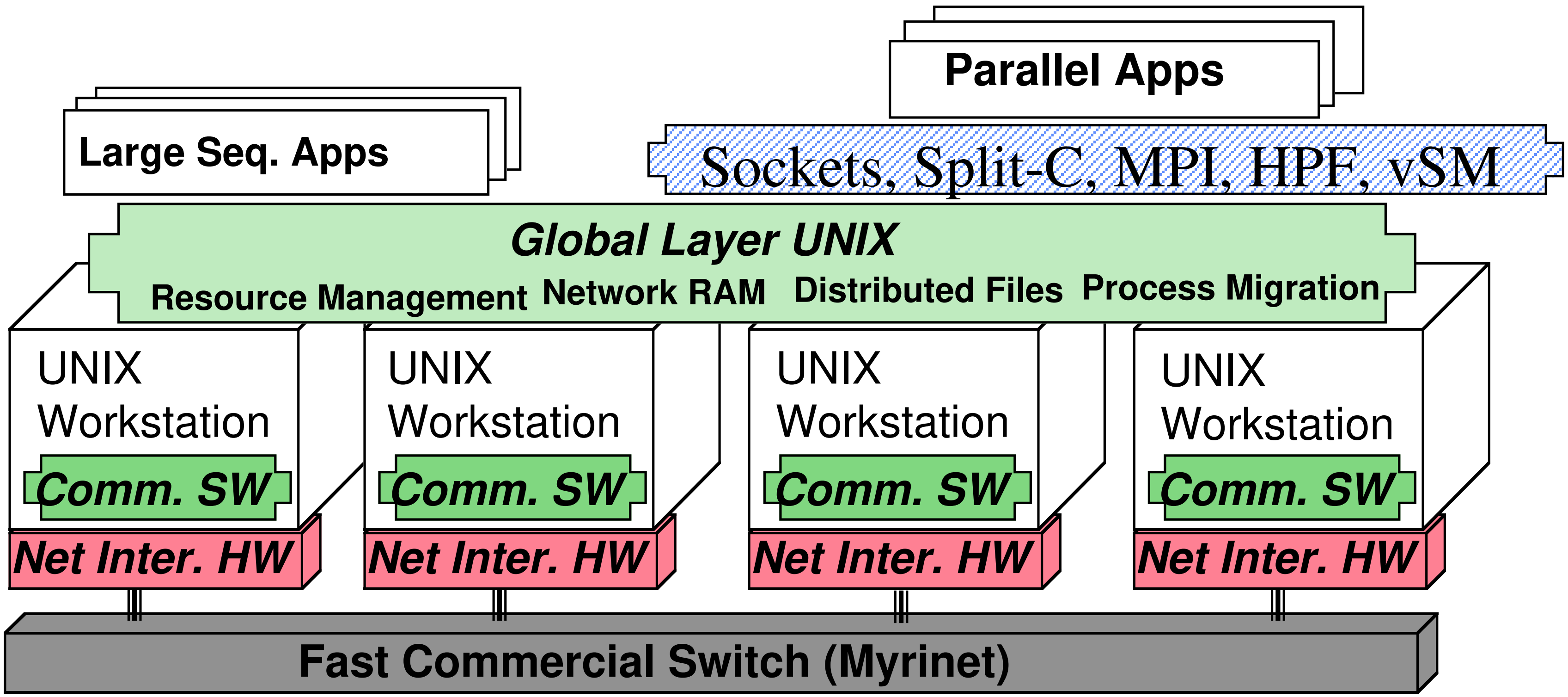
- Request / Reply small active messages (RPC)
- Bulk-Transfer (store & get)
- Highly optimized communication layer on a range of HW

Compare: Communication Time

Microseconds for 8-word Active Message



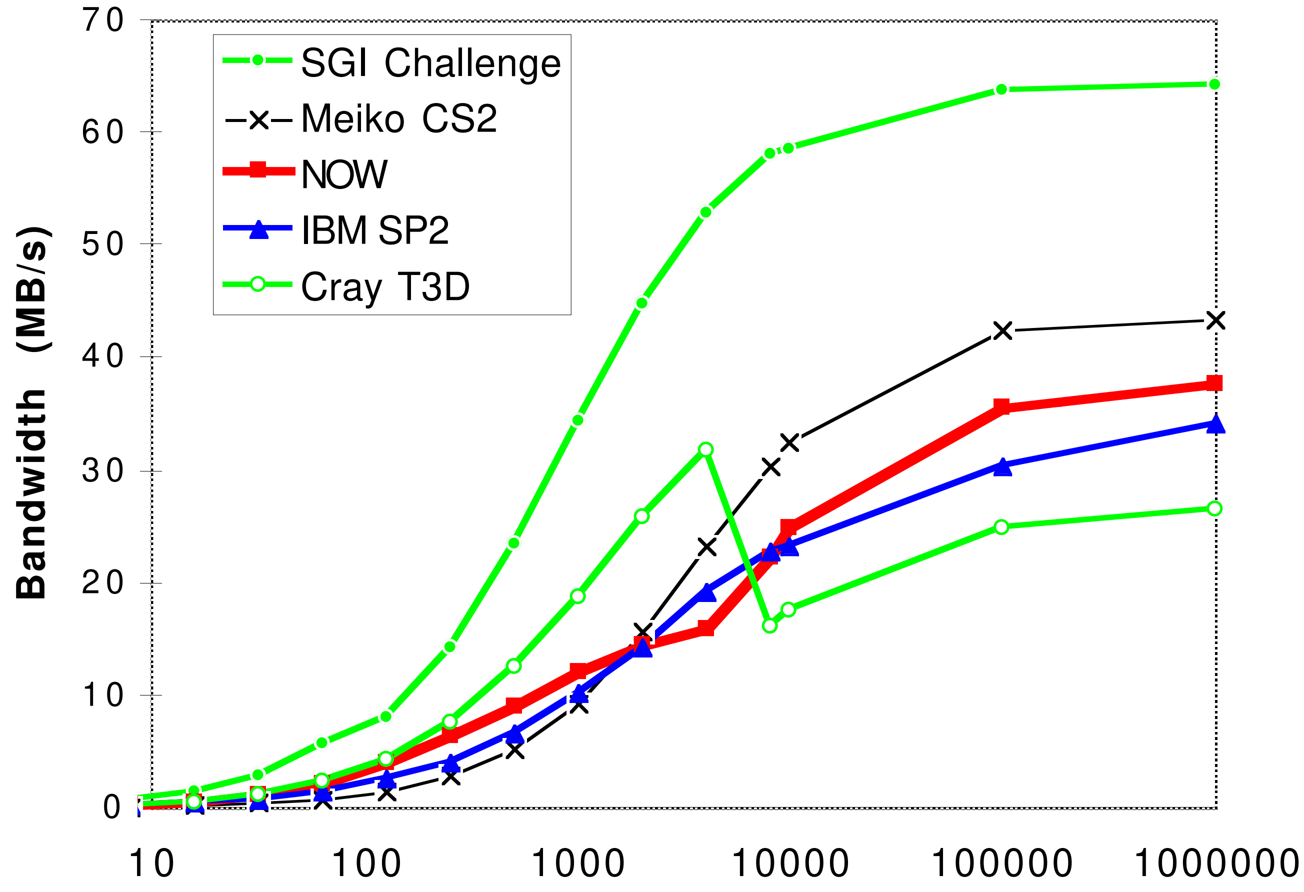
NOW System Architecture



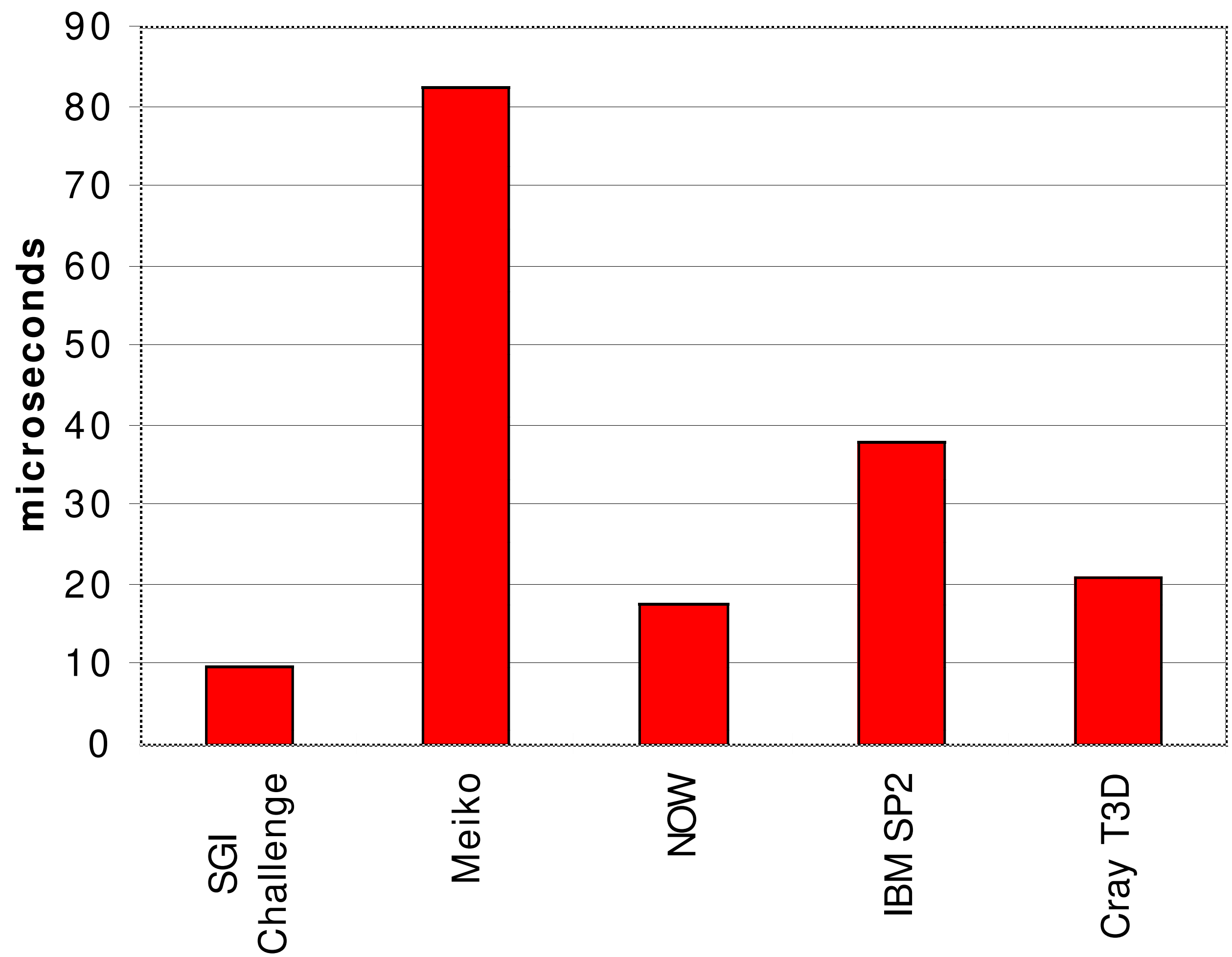
Approach: Build Traditional APIs on a light-weight protocol

- Compiler-based shared memory
- Shared virtual memory
- MPI
- HPF
- Sockets
- ...

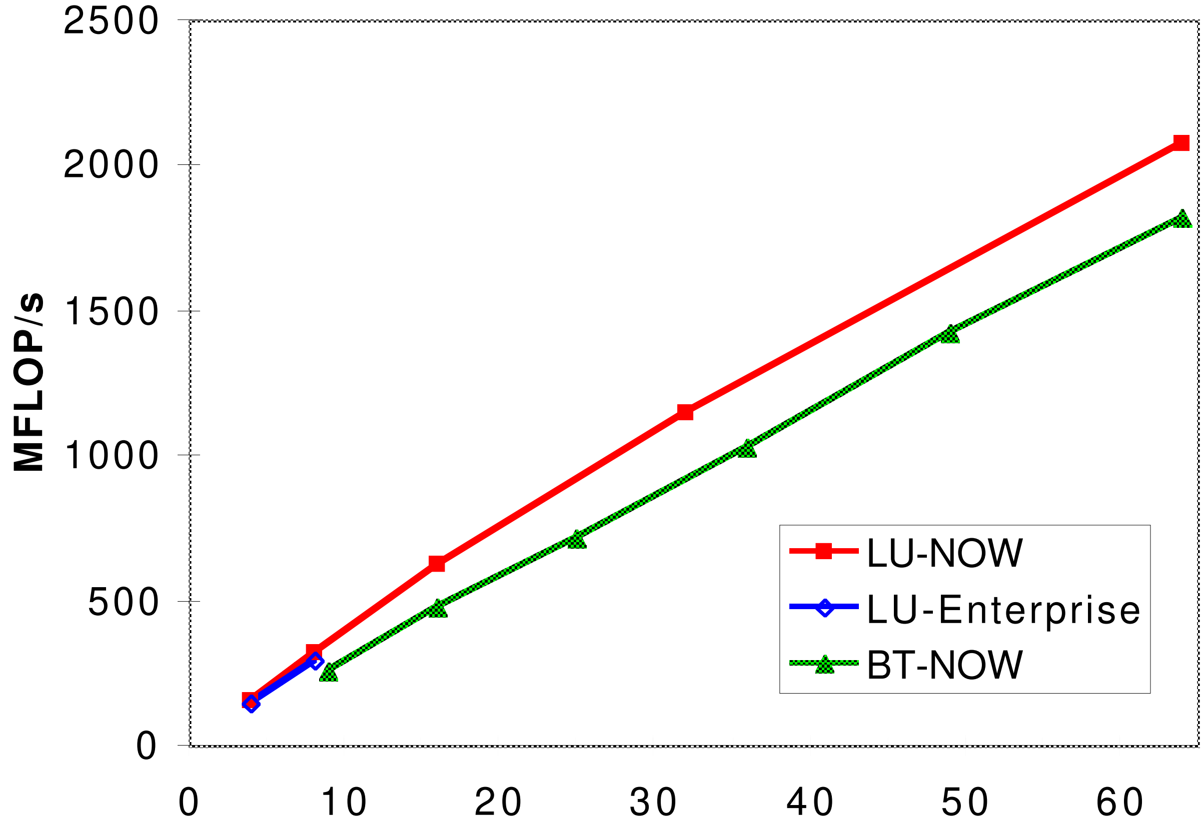
MPI over AM: ping-pong bandwidth



MPI over AM: start-up



NAS Parallel Banchmarks in MPI



What about sockets over AM?