



*Presentation to the IPDPS 2003 Workshop on  
Communication Architecture for Clusters:*

---

## ***Panel on Cluster Interconnect Technologies***

# **What are the Top 3 Limitations**

**Dr. Thomas Sterling**

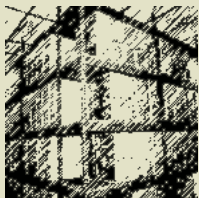
California Institute of Technology

and

NASA Jet Propulsion Laboratory

April 22, 2003

---



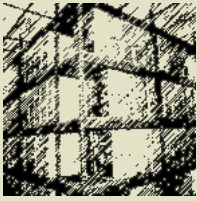
# LLNL MCR 7.6 Tflops





# Workload Classification with respect to Usage of Commodity Clusters

- ◆ Task stream well suited to commodity cluster
  - Regular coarse grain parallelism
  - Low communication and synchronization demands
  - Insensitive to latency, throughput oriented
  - Scales well
- ◆ Cluster low-cost compromise
  - Would run better on more tightly coupled machine
  - Cost matters – delivers best price-performance
  - Scalable across a reasonable regime
  - Note: some customers don't care about network cost
- ◆ Workload inappropriate for cluster solution
  - Medium to fine grain parallelism, possibly irregular
  - High communication demands and latency sensitive
  - Poor scalability



# Impact of Cluster Networks

- ◆ Integrate nodes
  - Gives rapid access for HPC to new technology
  - Kicker to small system scale
- ◆ Determine generality
  - Increases domain of problems that can run at least adequately
- ◆ Non-monotonic driver of performance to cost
  - While network is bottleneck, increases in BW can improve performance and initially performance to cost
  - When cost is too high, simply unavailable to some users



# List of priorities

- ◆ Don't worry
  - “Low hanging fruit opportunity”, not the ultimate solution
  - use clusters when they're good,
  - don't when they're not
- ◆ Cost
  - Really matters to many users, especially of moderate scale
  - Can be dominated by network, especially for large scale
- ◆ Scalability
  - Higher link bandwidth for more powerful nodes
    - SMPs
    - Bladed
  - Lower latency links and routers for more nodes
- ◆ Reliability
  - Stays up longer
  - Multiple paths for redundancy
  - Ease of maintenance



## But my real concern is:

---

- ◆ In the limit, clusters can only become as good as the best distributed memory COTS based MPPs
  - e.g. IBM SP-2, Intel Touchstone Paragon
- ◆ Redesign the microprocessor
  - Put system-wide communication in to the ISA for low overhead
  - Provide hardware support for strong latency hiding
    - e.g. multithreading, fine-grain transaction processing, percolation
- ◆ But ...
  - Then its not commodity
- ◆ Conclusion ...
  - If you're going to put a lot of work in to developing new networking hardware, why not just build a good computer instead.