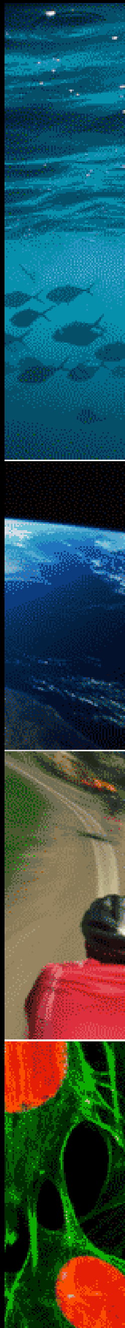
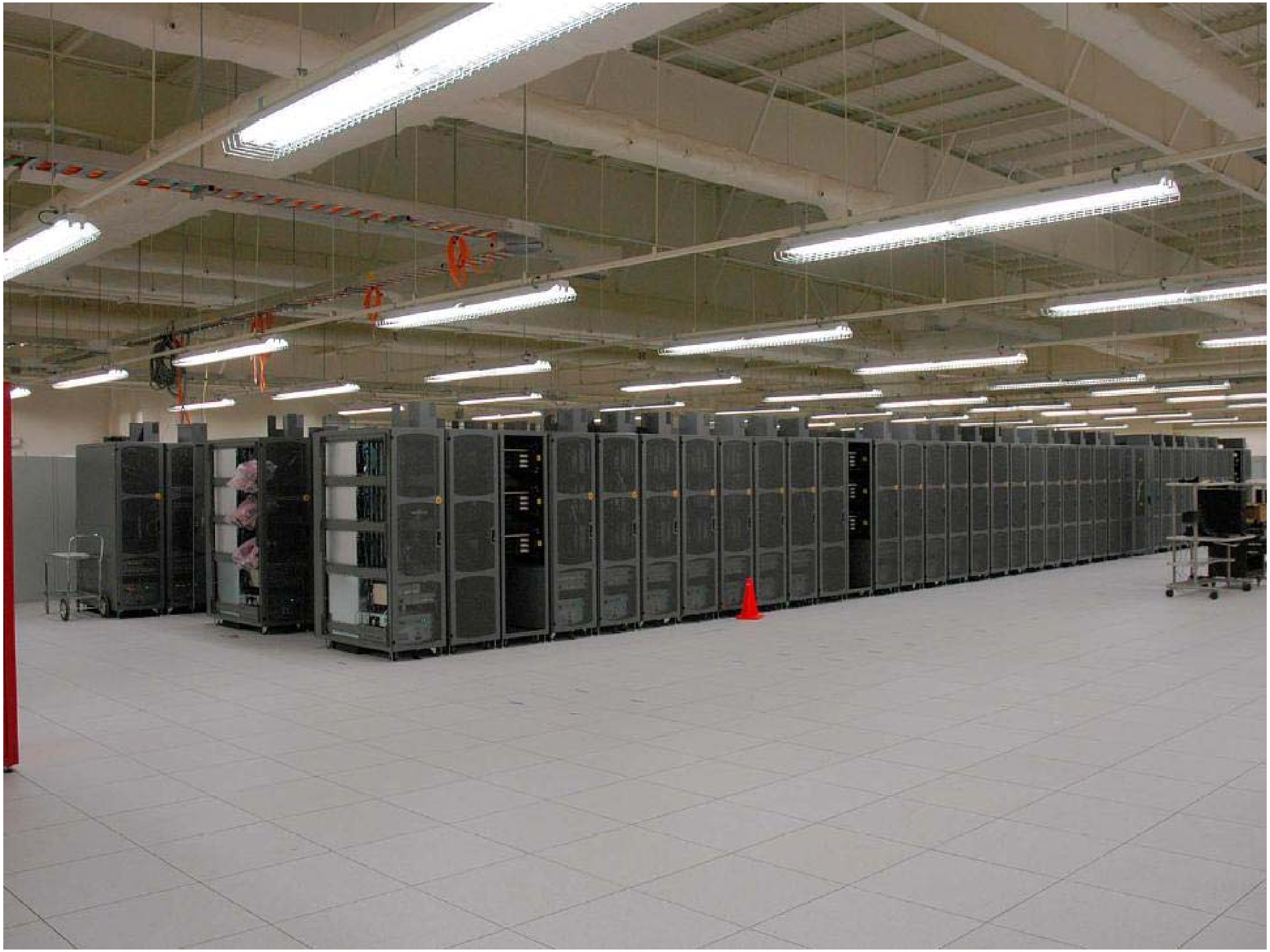
Cray High Speed Interconnect Requirements

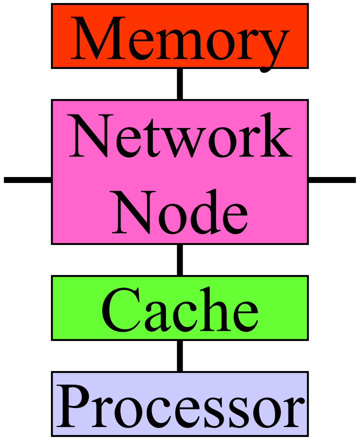
Mike Steinberger
Karim Tatah

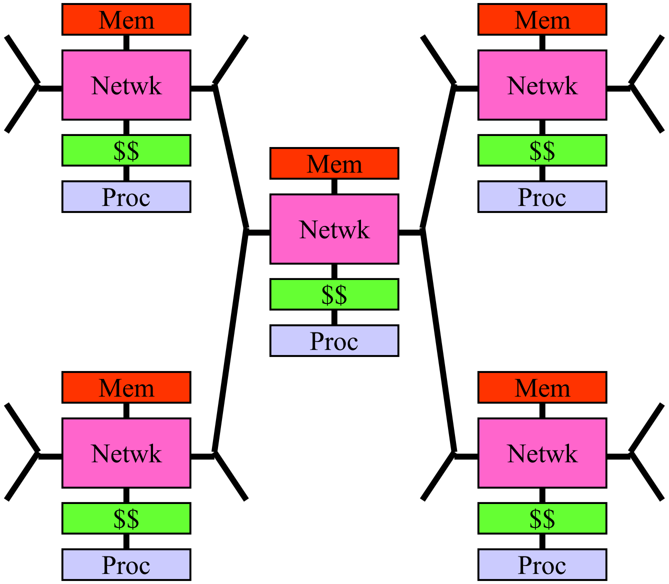


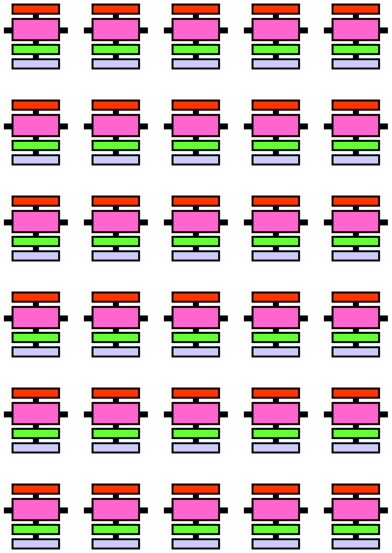
Typical Large Cray Supercomputer



Generic Supercomputer Architecture







As far as the budget extends…

•Each processor has access to all of the memory in the machine through an internal network with a global address space.

•Each network node is typically connected to many (e.g., six) other network nodes.

•Most systems extend across multiple chassis, each of which contains multiple PC boards, each of which contains multiple processors.

•Systems include I/O to an array of off the shelf peripherals such as disk drives.

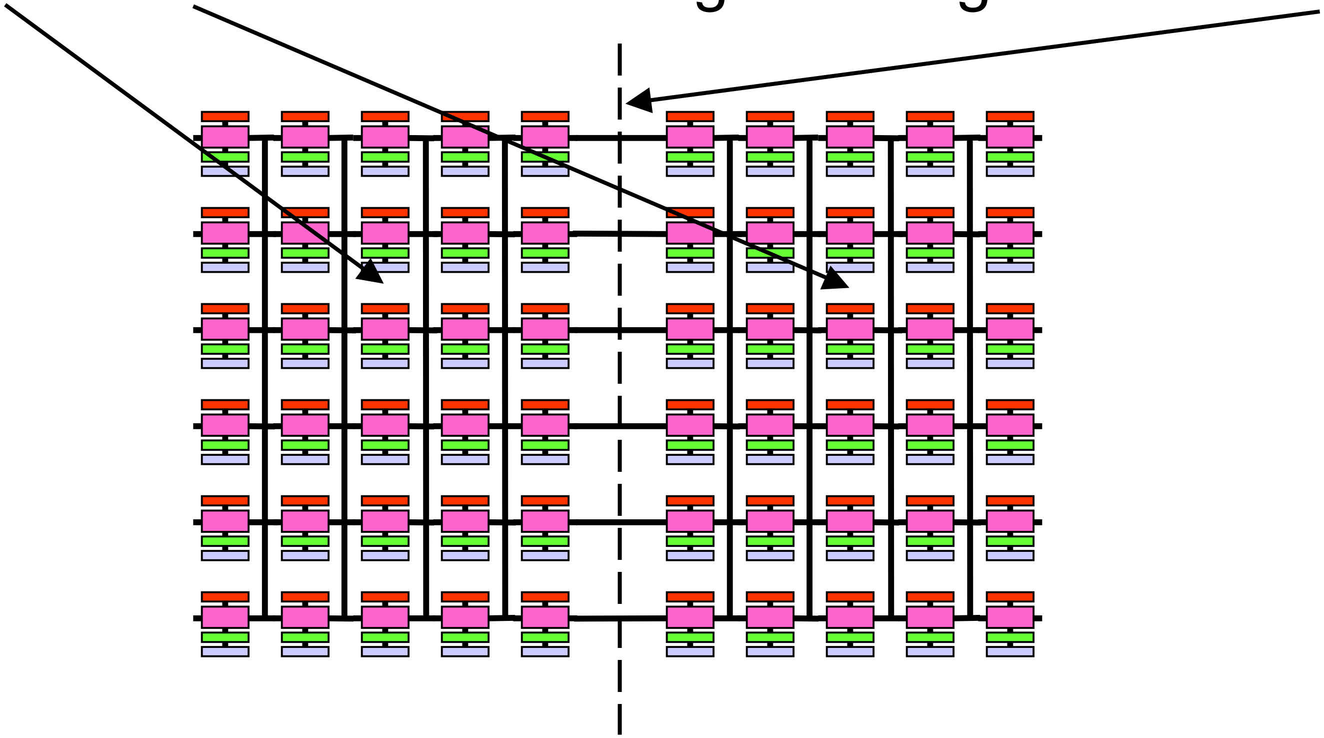
The Need for Speed (Channel speed, that is)

- On large, complex problems, processors must draw their operands from the entire memory address space.
- If processors have to wait for their operands, then the computational throughput is determined by network bandwidth, and not processor speed or the number of processors.
- Architectures for hiding memory subsystem latency
  - Vector: Perform the same operation on a number of operands in succession. Perform that operation on the first operands while the rest are still being fetched.
  - Multi-Threaded: While waiting for the operands for one task, work on another task for which the operands are already available.
- There is no way to hide a lack of bandwidth.

Network Bandwidth = Performance

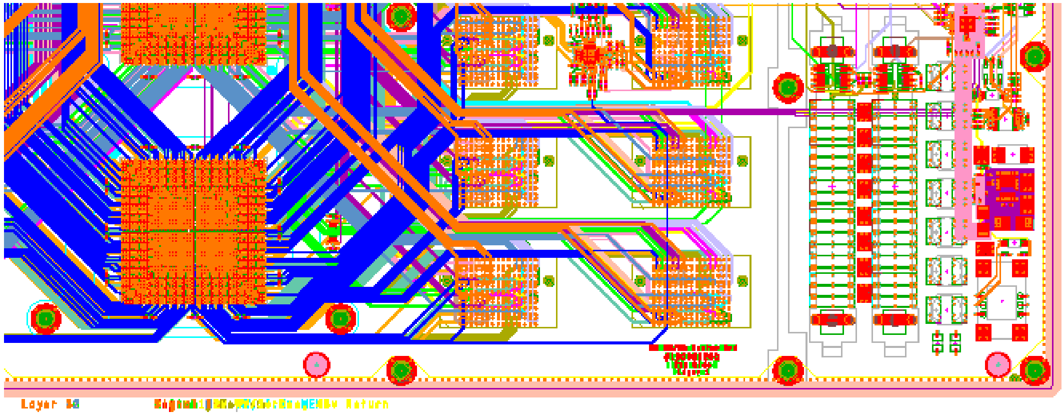
Bisection Bandwidth

It doesn’t matter much how fast data gets around here or here if it takes a long time to get across here.



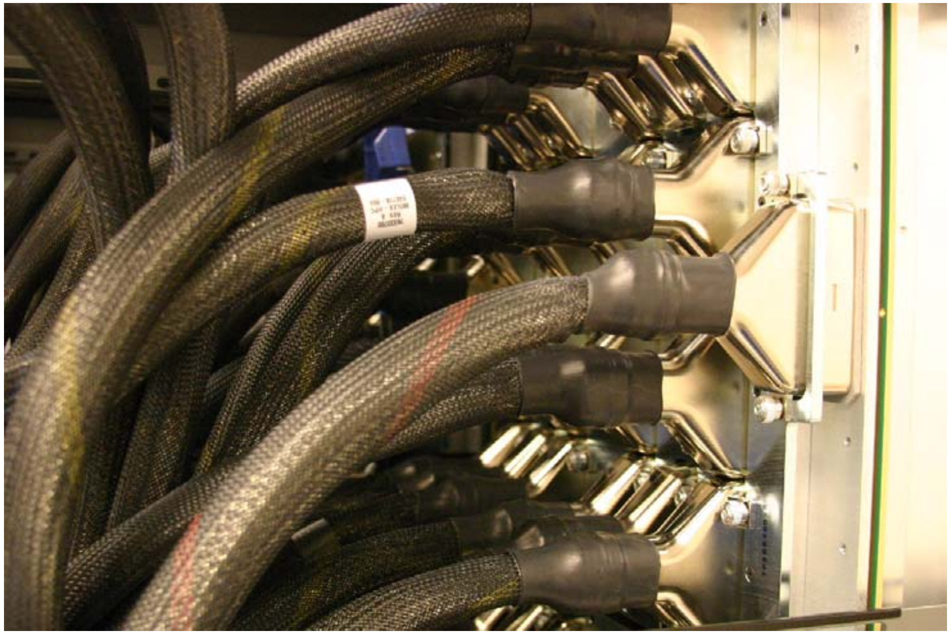
Performance is dominated by the lowest bandwidth path.

High Density Interconnect

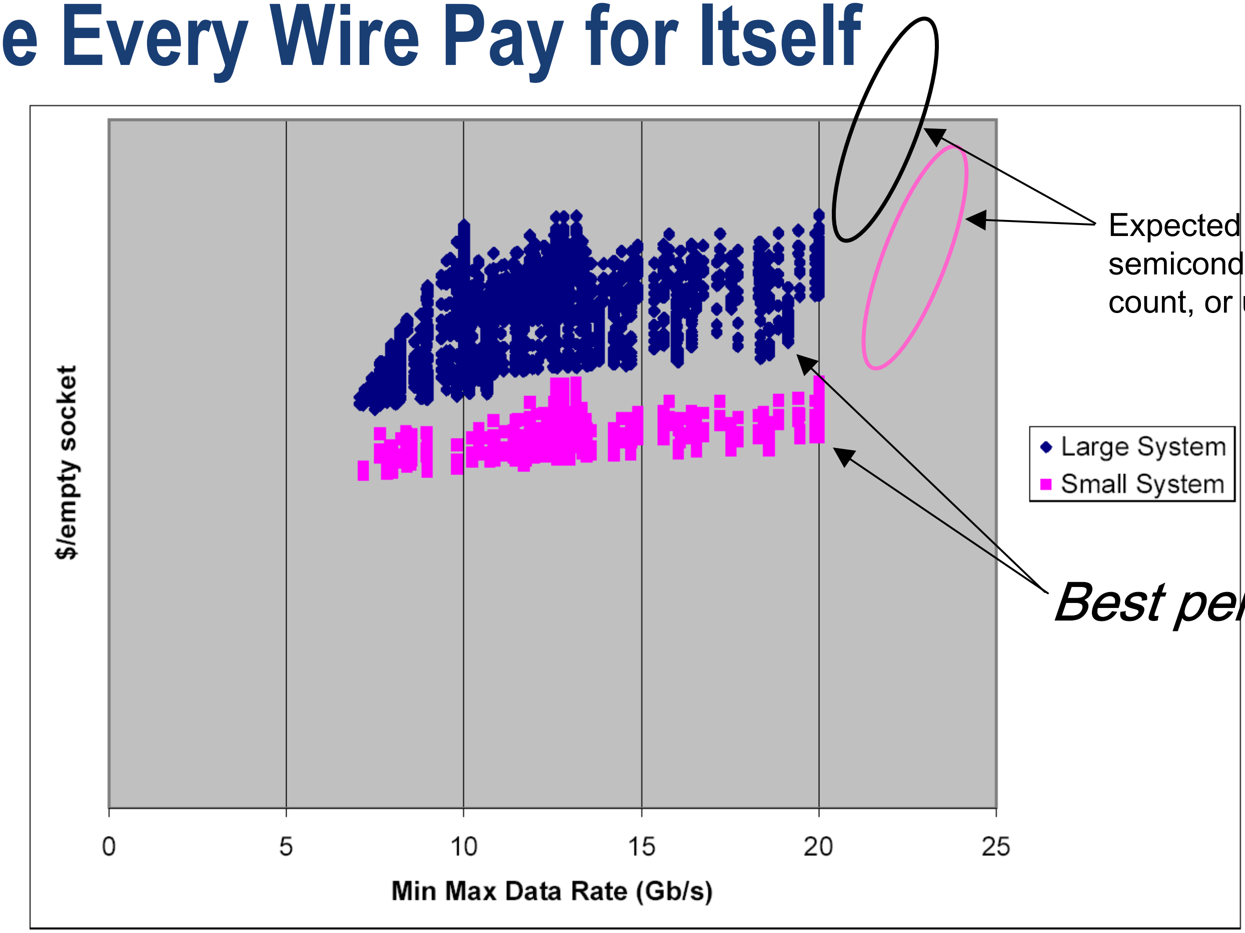


- >6 Gb/s Data Rate
- >1500 Differential Pairs
- 16 Signal Layers
- 5 Tb/s Tx, 5 Tb/s Rx

- 96 Differential Pairs per Cable



Make Every Wire Pay for Itself



Expected impact of compound semiconductors, increased parts count, or use of exotic technologies

Best performance/cost

- Two system configurations, each held constant. (I.e., PC board counts, interconnection distances)
- Every variation of the large system includes optics for the longest paths.
- Assumes CMOS pin electronics.
- Iterated over design choices that included variations in cable type, PC board material, use of electrical repeaters, use of optics.
- For each set of design choices, plotted empty socket cost (processor and memory cost excluded) and maximum data rate for worst case path.

Cray HSSG Requirements

- Optimized Physical Layer
  - Highest achievable data rate
  - Highest achievable bandwidth density
  - Minimum complexity
    - At these data rates, only simple solutions will be practical anyway.
    - Cray uses proprietary protocols for supercomputer internal networks.
  - At least 30 meters reach. 100 meters would be good.
- I/O Solution
  - Commercially supported interface to very large storage arrays and other peripherals
  - Very high bandwidth density