Some Thoughts On Interconnect Needs For ExaFlop Systems

Jim Tomkins

SOS13 Workshop
Hilton Head Island, South Carolina
March 9 - 12, 2009

Outline

- **Expectations**
- **A Few Things to Consider**
- **Some ExaFlop System Parameters**
- **Interconnect Performance Needs**

Expectations

- **Achieve a sustained ExaFlop on HPL - Probably will need a peak of ~1.5 ExaFlops**
- **System operation in ~2018**
- **System cost of $200M - $400M**
- **<40 MW (This will require some magic)**
- **Order of 100M Cores or more and 128K - 256K Processor Chips**
- **Want to achieve ~50% parallel efficiency across the whole machine (weak scaling) on important, scientific engineering problems of national interest**
- **Reliable enough to make significant progress between failures - >20 hours**

A Few Things to Consider - Programming Models

- **Message Size - Interconnect bandwidth balance will be even more difficult to achieve as nodes become more and more powerful**
  - **Three level data movement will probably be needed - memory, node, machine**
  - **Message overhead (hardware and software) will need to be minimized to optimize network delivered bandwidth**
  - **Message latency will be harder to hide as memory per core shrinks**
  - **Message throughput and bandwidth will probably need to increase faster than node performance because of smaller memories per core - ratio of computation to communication may be reduced which will put more stress on communication performance to achieve parallel efficiency**

A Few Things to Consider - Memory Bandwidth

- **Overlapping computation with communication is difficult to do for many applications. Even if possible, both computation and communication need memory bandwidth.**
- **Overlapping Communication with Communication can be done for most applications but both operations need memory bandwidth.**
- **Memory bandwidth balance (B/F) has been decreasing for years and is likely to continue to decrease even though the memory bandwidth per socket has been increasing.**
- **Memory bandwidth limits may limit network injection bandwidth**

A Few Things to Consider - System Topology

- **All topologies other than a single stage crossbar have contention in the network - Fat Trees, Meshes / Toruses, Hypercubes, Etc.**
- **Network link bandwidth needs to be higher than injection bandwidth to deal with contention**
- **The larger the system the greater the link bandwidth needed to deal with contention**
- **Network physical scalability for an ExaFlop system may drive cost and engineering considerations for the network and may be a major factor in determining topology for very large systems**

2018 Technology Assumptions

- **Will Moore's Law Continue to hold (8 nm in 2018) - Will cost of Fabs end it?**
- **Multi-Core Chips - Still More Cores**
  - **Clock Rate Increases are stalled - power considerations will probably push clock rates lower to optimize performance per Watt**
  - **Greater Parallelism in processor chips - 128 - 1024 cores per chip (maybe more)**
  - **Will lack of commercial viability for very high core count chips end the rate of increase?**
- **Memory System**
  - **Memory per Processor Core will Decrease - Cost and Power**
  - **Memory per Op will continue to Decrease**
  - **Memory Bandwidth - WDM Direct Optics off Chip should boost Bandwidth, but will need a large number of memory Banks to reduce contention**
  - **Latency is Constant at best - probably will increase some**
  - **Soft Errors will become even more of a problem - Particularly for functional units**

2018 Technology Assumptions

- **System Interconnect Performance**
  - **Bandwidth - WDM Direct Optics off Chip should boost Bandwidth**
  - **Absolute Latency is slowly Decreasing - Need even smarter NICs**
  - **Message Throughput could still be Increasing with Smarter NICs**

Possible 2018 Systems?

	System 1	System 2	System 3
Peak ExaFlops	~1.5	~1.5	~1.5
Sockets	131,072	65,536	262,144
Cores/Socket	512	128	1024
Clock Speed (GHz)	1.4	2.8	1.4
Socket Peak (GF)	11,468.8	11,468.8	5,734.4
Memory B/W (B/F)	0.1 - 0.5	0.1 - 0.5	0.1 - 0.5
Memory per Socket (GB)	256 - 512 (0.5 - 1.0 GB/core)	512 - 1024 (4.0 - 8.0 GB/core)	128 - 256 (0.125 - 0.25 GB/core)
Link B/W (B/F)	0.5 - 1.0	0.5 - 1.0	0.5 - 1.0
System Power (MW)	60 - 80 (<40)	80 - 120 (<40)	60 - 80 (<40)
Floor Space (sq ft)	~12,000	~8,000	~18,000
Programming Model	Multi-Level Explicit Message Passing?	Multi-Level Explicit Message Passing?	Multi-Level Explicit Message Passing?

COTS Processor On Chip Interconnects

- Internal Chip Topology in 2018
  - Meshes
  - Fat Trees
- Issue - How to maintain internal Bandwidth and Latency balance as the number of cores increases dramatically
  - Latency will grow
  - Bandwidth will probably decrease on a per core basis

System Interconnects

	System 1	System 2	System 3
System Size Sockets Peak ExaFlops TF/Socket	 131,072 ~1.5 11.4688	 65,536 ~1.5 11.4688	 262,144 ~1.5 5.7344
NIC B/W (B/F)	0.25 - 0.5	0.25 - 0.5	0.25 - 0.5
Link B/W (B/F)	0.5 - 1.0	0.5 - 1.0	0.5 - 1.0
MPI Latency (ns)	<500	<500	<500
MPI Throughput (M Msg/s/NIC)	>1000	>1000	>500
Load/Store (M Msg/s/NIC)	>5000	>5000	>2500
Load/Store Latency (ns)	<150	<150	<150

Final Thoughts

- **Will Moore's Law continue for 6 more generations? The technology is there but is the economic case there?**
- **The level of parallelism is growing rapidly with the growth of the number of cores per chip and wider functional units. Will commercial applications take advantage of all of the parallelism and provide the economic incentive to the processor companies to build the high core count chips?**
- **Scalability for science and engineering codes will depend on paying attention to data locality. Will programming models make that possible?**
- **Performance per socket for science and engineering codes is likely to continue to become less well balanced as system peak grows. We will get a smaller and smaller fraction of the peak unless we can convince the computer companies to pay more attention to getting more data on and off of the processor chip.**

Final Thoughts

- **Parallel efficiency is likely to dominate overall system performance at ExaFlop scale. System interconnect performance is the most important hardware contributor to parallel efficiency. It is likely to become even more important to parallel efficiency as the number of cores continues to rapidly increase and the memory per core decreases.**
- **Maintaining even the current level of system balance is going to be very difficult between now and 2018.**
- **Will an ExaFlop system be practical in 2018 (power, scalability, reliability)?**