Interconnects

A Buyers Point of View

Jim Tomkins

ACS Workshop
Baltimore, MD
June 13 - 14, 2007

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

2011 COTS Technology Trends

- Moore’s Law - Transistor density is still on the curve (32 nm)
- Multi-Core Chips
  - Clock Speed Increases Stalled
  - Wider Functional Units - 8 to 16 Flops/clock
  - Greater Parallelism - 4, 8, 16, 32 cores
  - Large General Purpose Systems may have a million or more cores by 2011
- Memory System
  - Memory per Processor Core is Decreasing
  - B/F Ratio is Decreasing
  - Latency is Constant or Increasing
- System Interconnect Performance
  - B/F Ratio is Decreasing - SERDES speed and pin count are limiting
  - Absolute latency is Decreasing - Smart NICs

2015 COTS Technology Trends

- **Moore's Law Continues to hold (15 nm)**
- **Multi-Core Chips - More Cores**
  - **Clock Rate Increases still stalled**
  - **Greater Parallelism in functional units - maybe vectors**
  - **Greater Parallelism - 32 - 128 cores**
  - **Large General Purpose Systems may have 4 million or more cores**
- **Memory System**
  - **Memory per Processor Core is probably Decreasing**
  - **Memory per Op is Decreasing**
  - **Memory Bandwidth - Capacitive Coupling? Direct Optics off Chip? - Could stop slide in B/F ratio, memory Banks per DiMM**
  - **Latency is Constant or Increasing**
- **System Interconnect Performance**
  - **Bandwidth - Direct Optics off Chip will be needed to maintain B/F ratio**
  - **Absolute Latency is slowly Decreasing - Smarter NICs could help**

2019 COTS Technology Trends

- Moore’s Law Continues to hold (8 nm) - Will cost of Fabs end it?
- Multi-Core Chips - Still More Cores
  - Clock Rate Increases still stalled - New Technology?
  - Greater Parallelism in functional units - more vector pipes
  - Greater Parallelism - 128 - 512 cores
  - Large General Purpose Systems may have 16 million or more cores
- Memory System
  - Memory per Processor Core is probably Decreasing
  - Memory per Op is Decreasing
  - Memory Bandwidth - WDM Direct Optics off Chip should boost Bandwidth, memory Banks per DiMM
  - Latency is Constant at best
- System Interconnect Performance
  - Bandwidth - WDM Direct Optics off Chip should boost Bandwidth
  - Absolute Latency is slowly Decreasing - Need even smarter NICs

Sandia National Laboratories

Possible 2011 Systems

	System 1	System 2
Peak PF	5 - 6	15 - 32
Sockets	32,768	32,768
Cores/Socket	8	16 - 32
Clock Speed (GHz)	2.4 - 3.0	2.4 - 3.0
Socket Peak (GF)	150 - 190	600 - 1000
Memory B/W (B/F)	~0.25	~0.1
Memory per Socket (GB)	16 - 32	16 - 32
Link B/W (B/F)	0.04 - 0.5	0.01 - 0.1
System Power (MW)	7 - 10	10 - 15
Floor Space (sq ft)	~5000	~5000
Programming Model	Explicit Message Passing	Explicit Message Passing

Possible 2015 Systems

	System 1	System 2
Peak PF	20 - 50	80 - 200
Sockets	32,768	32,768
Cores/Socket	32	64 - 128
Clock Speed (GHz)	2.4 - 3.0	2.4 - 3.0
Socket Peak (GF)	600 - 1500	2400 - 6100
Memory B/W (B/F)	0.05 - 0.1 (1.0)	0.025 - 0.05 (1.0)
Memory per Socket (GB)	64 - 128	64 - 128
Link B/W (B/F)	0.02 - 0.12 (0.1 - 1.0)	0.005 - 0.03 (0.025 - 0.25)
System Power (MW)	9 - 12	15 - 20
Floor Space (sq ft)	~5000	~5000
Programming Model	Explicit Message Passing?	Explicit Message Passing?

Possible 2019 Systems

	System 1	System 2
Peak PF	80 - 200	325 - 820
Sockets	32,768	32,768
Cores/Socket	128	256 - 512
Clock Speed (GHz)	2.4 - 3.0	2.4 - 3.0
Socket Peak (GF)	2,400 - 6,000	10,000 - 25,000
Memory B/W (B/F)	0.1 - 1.0	0.025 - 0.25
Memory per Socket (GB)	256 - 512	256 - 512
Link B/W (B/F)	0.1 - 1.0	0.025 - 0.25
System Power (MW)	12 - 15	20 - 25
Floor Space (sq ft)	~5000	~5000
Programming Model	Explicit Message Passing?	Explicit Message Passing?

COTS Processor On Chip Interconnects

- Internal Chip Topology
  - Small numbers of Cores - Cross Bars
  - Small to medium numbers of Cores - Rings
  - Medium to large numbers of Cores - Meshes, Fat Trees
- 2011: small to medium
- 2015: medium to large
- 2019: Large
- Issue - How to maintain internal Bandwidth and Latency balance as the number of cores increases dramatically

System Interconnects

	2011		2015		2019	
System Size						
Sockets	32,768		32,768		32,768	
Peak PF	32		200		800	
TF/Socket	1.0		6.1		25.0	
	Expect	Want	Expect	Want	Expect	Want
NIC B/W (B/F)	0.01 - 0.1	1.0	0.005 - 0.03	1.0	0.025 - 0.25	1.0
Link B/W (B/F)	0.01 - 0.1	1.0	0.005 - 0.03	1.0	0.025 - 0.25	1.0
MPI Latency (ns)	750 - 1500	500	500 - 1000	400	400 - 750	300
MPI Throughput (M Msg/s)	20	50	80	300	300	1200
Load/Store (M Msg/s)	75	400	150	1,600	300	6400
Load/Store Latency (ns)	300	100	300	100	300	100

Final Thoughts

- Will Moore's Law continue for 6 more generations?
- The level of parallelism is growing rapidly with the growth of the number of cores per chip and wider functional units.
- Scalability will depend on paying attention to data locality.
- Performance per socket 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.
- System interconnect performance will become even more important to overall system performance 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 2019.