Panel : Issues for Exascale Scalability and Resilience

OPENFABRICS ALLIANCE

Panel Members:

Matt Leininger

Ronald Luijten

Steve Poole

Ishai Rabinovitz

Bob Woodruff

Susan Coulter

Systems	2009	2015 +0/-1	2018 +1/-1
System peak	**2 Peta**	**100-200 Peta**	**1 Exa**
Power	**6 MW**	**~15 MW**	**~20 MW**
System memory	0.3 PB	5 PB	64 PB (+)
Node performance	125 GF	0.5 TF or 7 TF	1,2 or 15TF
Node memory BW	25 GB/s	1-2TB/s	2-4TB/s
Node concurrency	12	O(100)	O(1k) or 10k
Total Node Interconnect BW	3.5 GB/s	100-200 GB/s 10:1 vs memory bandwidth 2:1 alternative	200-400GB/s (1:4 or 1:8 from memory BW)
System size (nodes)	18,700	50,000 or 500,000	O(100,000) or O(1M)
Total concurrency	225,000	O(100,000,000) *O(10)-O(50) to hide latency	O(billion) * O(10) to O(100) for latency hiding
Storage	15 PB	150 PB	500-1000 PB (>10x system memory is min)
IO	0.2 TB	10 TB/s	60 TB/s (how long to drain the machine)
MTTI	days	O(1day)	O(0.1 day)

Some things we are missing (IMHO)

- Adaptive /Dynamic Routing *
- Congestion Control *
- PGAS elements (SW/HW/FW)
- A memory interface design
  - Generic (QPI, HT*,Power*,DDR*…)
  - Busses are WAY TOO slow
- Wider number of lanes. (i.e.16/32/64) *
- Atomic Memory Operations *
  - Store, fetch&add, compare&swap, put, get, barrier…

Some things we are missing (IMHO)

- Collectives (FP, Int, Logical) (prefer HW) *
  - All-reduce, add, sum, min, max…
- IB/Ethernet

  (It is faster for IB to build a 100GE NIC than others)
- More aggressive optics (links, *photonics…)
- More aggressive power savings elements
  - We can not tolerate a 50-100MW ExaOp machine
- High Message Injection rate (small messages) *

Some things we are missing (IMHO)

- COE *
- Reliable Multicast *
- Better Fiber ?
  - Better BER, everywhere *
- Remember File I/O
- Remember Stack/code bloat ;-)
- Remember “remote”
  - Network + File I/O + PB/EB
- Need 1/10/100Tb networks (dates)