IBM Research

Challenges of Sustained PetaScale Performance

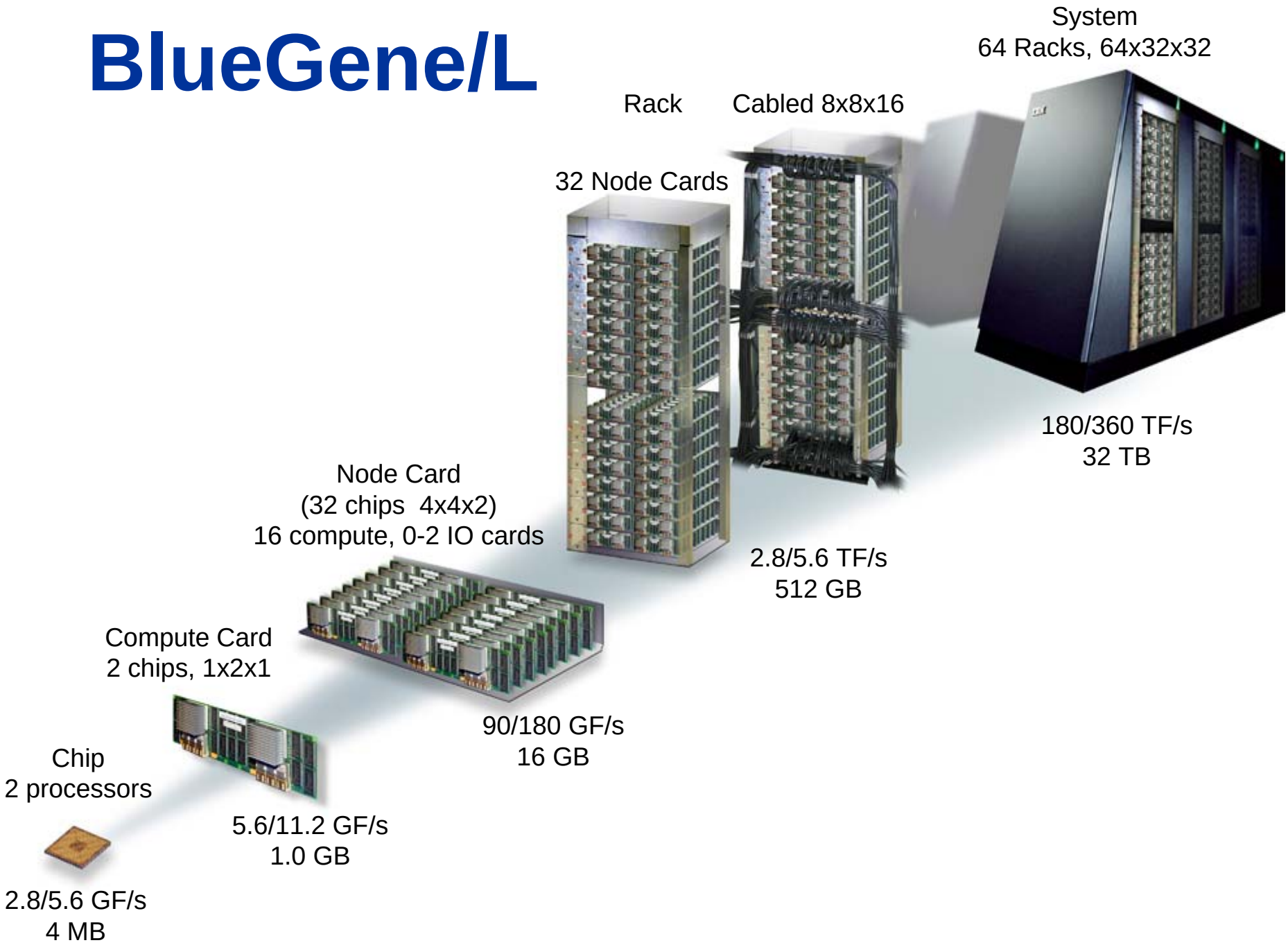
Multi-cores, accelerators, memory latency, memory capacity, memory/network/interconnect technologies, **packaging**…

Paul Coteus, IBM

6/12/2007

© 2007 IBM Corporation

BlueGene/L


System
64 Racks, 64x32x32
Rack
Cabled 8x8x16
32 Node Cards
180/360 TF/s
32 TB
Node Card
(32 chips 4x4x2)
16 compute, 0-2 IO cards
2.8/5.6 TF/s
512 GB
Compute Card
2 chips, 1x2x1
90/180 GF/s
16 GB
Chip
2 processors
5.6/11.2 GF/s
1.0 GB
2.8/5.6 GF/s
4 MB

Multi-PetaFLOP system choices, in order of importance

- What is our power budget?
- What is our reliability target? (Soft errors becoming very important!)
- What is our interconnect? (commercial/custom, electrical/optical, …)
- What are our memory system metrics?
- What is our processor? (monolithic/embedded, multi-threaded, …)

(Probably should have put cost first, but it factors into all items)

Note: The following scaling is pure conjecture on my part, and is not to be construed as any product plan, etc.. I apologize in advance for any blatant errors.

© 2007 IBM Corporation

Scaling "1 PetaFLOP" to 1 ExaFLOP – Power & Reliability

From PetaFLOP to ExaFLOP	BG/L	Mult	1000 PetaFLOP	Comment on ExaFLOP system
Target power dissipated (Compute, MW)	**1.5**	21.6	32	
# of racks	**64**	6.0	384	
Power / rack (KW)	23		83	**Water Cooling**
(Electrically connected) midplanes per rack	**2**	0.5	1	
Midplanes per system	128		384	
Target MTBF (days)	**5**	**1.0**	5	
Failure rate of processor node (x10-9/hr, node is 80%)	**106**	**1.3**	141	**SER control!**
Number of processor nodes	65,536		49,152	
Processor stacks per node	**1**	**1.0**	1	
Processor stacks per midplane	512		128	
Power_proc / Power_mem	**2**	0.5	1	
Power of a processor chip stack (W - node is 75%)	11		242	**Thermal management**

1ExaFLOP with 5 day MTBF and 32 GigaWatt dissipated power =>
384 racks @ 83 KW each. 128 nodes/rack, node is ~242W processor-stack

© 2007 IBM Corporation

Scaling "1 PetaFLOP" to 1 ExaFLOP
– Stacking, Multi-core, Accelerators, Frequency

From PetaFLOP to ExaFLOP	BG/L	Mult	1000 PetaFLOP	Comment on ExaFLOP system
PEAK FLOP Rate				
Stacking: Die per stack	1	6.0	6	
Multi-core: Cores per die	2	32.0	64	
Accelerators: FLOPs per cycle per core	4	4.0	16	
Frequency of core (GHz)	0.7	4.7	3.3	
Peak FLOP rate, relative to last generation		**2725**		
Peak FLOP rate (PetaFLOPs)	**0.37**		**1000**	**3D required**

The hypothetical ~240W processor chip-stack has 6 stacked chips, each with 64 sets of 3.3GHz core + 8-way Floating Multiply-Add unit.

© 2007 IBM Corporation

Scaling "1PetaFLOP" to ExaFLOP - Network

From PetaFLOP to ExaFLOP	BG/L	Mult	1000 PetaFLOP	Comment on ExaFLOP system
NETWORK				
Electrical signaling rate (Gb/s)	**1.4**	5.7	8	
Electrical Ports (B)	**3**	10.7	32	
Optical signaling rate (Gb/s)		4.0	40	
Optical Ports (B)		32.0	32	
Network GB/s/stack	4		1280	**~1TB/s**
B/FLOP/core Network	**1**		**0.06**	**Not BiSection Bandwidth**

An "electrical" interconnect network, 32B@8Gb/s + an "optical" interconnect network, 32B@40Gb/s signaling results in over 1 TeraByte/s network bandwidth. Depending on topology, it may be acceptable.

© 2007 IBM Corporation

Scaling "1PetaFLOP" to ExaFLOP - Memory

From PetaFLOP to ExaFLOP	BG/L	Mult	1000 PetaFLOP	Comment on ExaFLOP system
MEMORY SYSTEM				
L2 per die (MB)	4	32.0	128	
MB L2 per core	**2**		**2**	
Memory data rate (Gb/s)	**0.350**	15.2	5.3	**DDR5 not fast enough!**
Memory channels	**1**	16.0	16	
Memory channel (B)	**16**	2.0	32	
Mainstore Bandwidth (GB/s)	5.6		2724	
B/Flop/core Bandwidth	**1**		**0.13**	
GB/node (SMP stack)	**1**	256.0	256	
Gb/DRAM	**0.5**	16.0	8	**8Gb not dense enough!**
Data DRAMs per chip	16		256	
Width of DRAM (B)	1		2	
Power per DRAM (W)	0.28	DDR5	0.76	
Mem capacity B/F	**0.2**		**0.013**	

256 8Gb/s DRAMs, each 2B wide and ~.75W, signaling at 5Gb/s, yield 2.7TB/s. 7x less B/F bandwidth, 14x less B/F capacity than BG/L. Need new technology. Also how to load/unload 12PB?

© 2007 IBM Corporation

Conclusions

- 1 PetaFLOP sustained (~10PF peak) is possible by end of the decade using aggressive embedded technology, mixed electrical-optical network, and standard high speed DRAM.
- Packaging considerations (power, failure rate, signaling speeds, etc dictate most of the properties of the system.
- 3D integration can yield 1 ExaFLOP peak system with <50MW, if soft errors can be controlled.
- Maintaining network B/F may be possible with multiple 40GB/s optical channels.
- However, maintaining external memory system B/F is not possible, even with x16 5Gb/s SDRAM-DDR5. Requires new (embedded?) memory technology.
- Loading and unloading the external memory system from file system is an even more difficult problem. Could both problems be solved with a new very dense, non-volatile, persistent memory?

© 2007 IBM Corporation