The silver lining of future systems with extreme concurrency: Node cost reductions will meet or exceed historic trends

Alan Gara
IBM Research

The Promise

Hippocratic Oath for Production System Supercomputer Architects

First , do no harm…

With little or no effort on the users part, performance on next generation production systems should achieve same or better performance of past machines of similar price.

The box we have traditionally found ourselves in…

The Memory per Core or memory per thread constraint severely restricts our possible directions and opportunities to provide higher perf/$

Flat MPI program model put system design in a box

- Programs based on MPI.
- MPI tasks have a minimum memory requirement typically between 250 MB to 4 GB.
- Memory was the most expensive system component
- Results in…
  - Power efficiency gains are difficult due to the pressure to have strong single thread. (compromises were necessary..)
  - Cost improvements bounded by memory size constraint.

The silicon trade off…. (mid 2011 time frame)

Compute size (12.8 GF floating point unit) 1.53mm^2

DRAM chip (4 Gb) 125 mm^2 30 Mb/mm^2

For 16 GB/chip in mid 2011 ratio…

Processor FPU area = 24 mm^2

16 GB of DRAM = 4800 mm^2

•There are significant differences in yield and process but the ratio of silicon area is ~ 200x (even with FPU = 1/10 compute area it is still 20x)

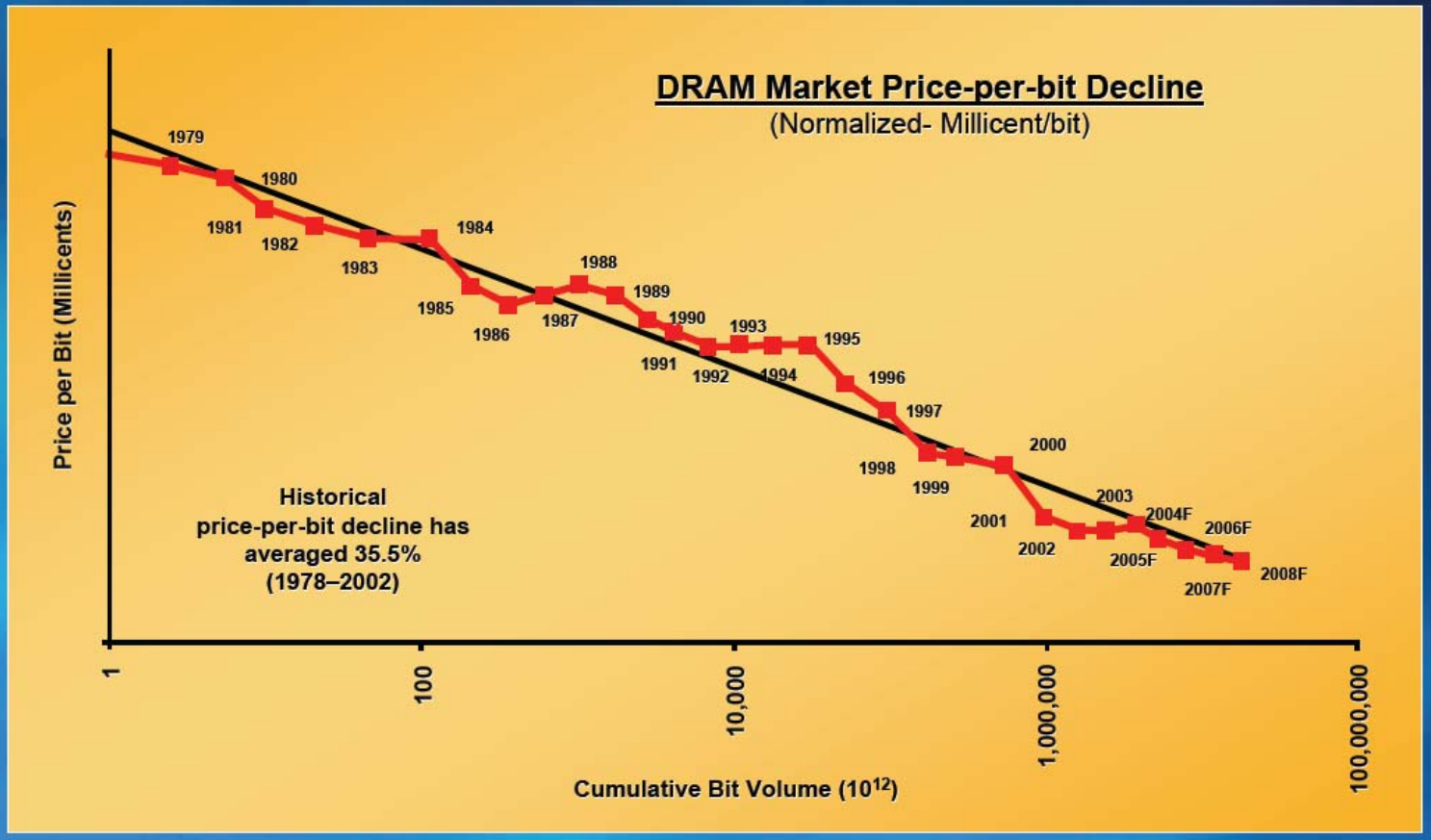
•DRAM price improvement is highly correlated to feature size hence doubling every two years is all that can be expected.

•Price/Performance of overall compute is not silicon dominated and can grow much faster…. if we enable it.

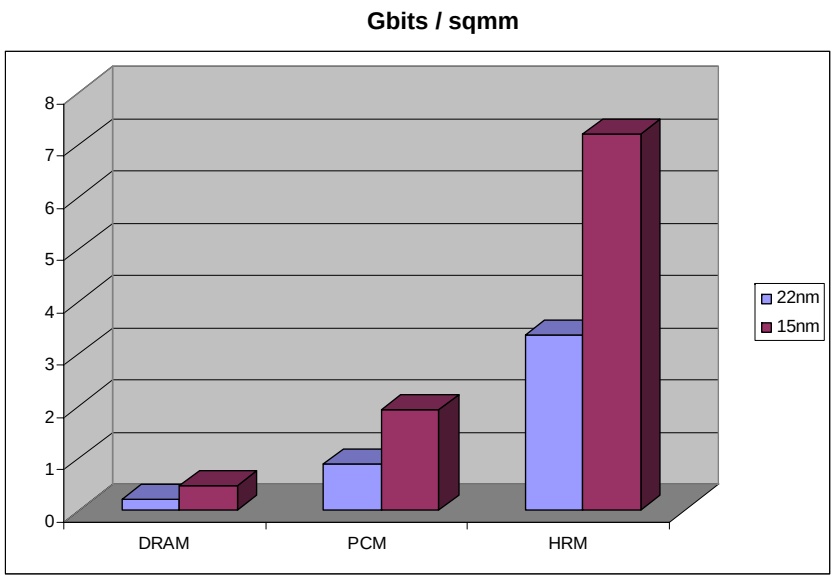
•Memory size to bandwidth ratio also locks this cost barrier in.

- ¼ to 1 GB/s per Gb is 2011 standard.
- 1 TB/s of memory bandwidth will require 160,000 mm^2 to 40,000 mm^2 of DRAM silicon if ratio holds from 2012 to 2015

Historical DRAM Price-Per-Bit Decline ~35%/year



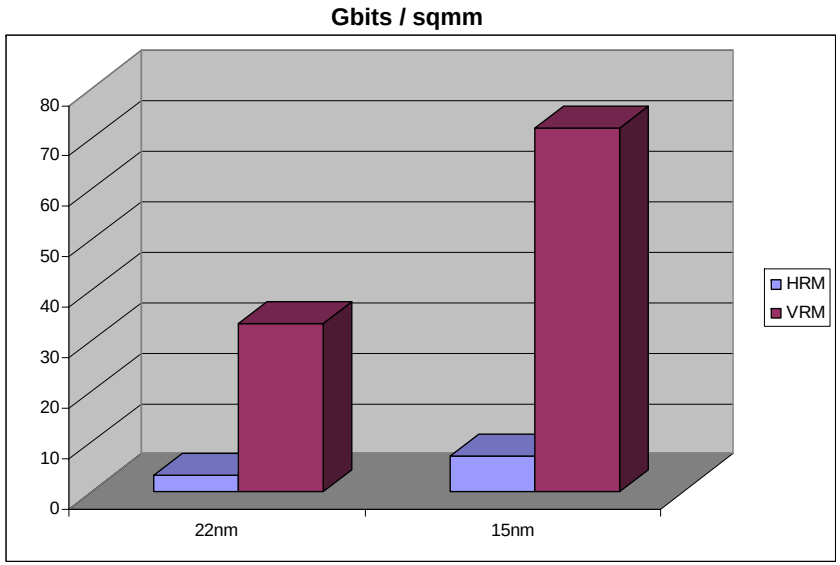
Memory Technology Density



Most activities targeted towards storage

- **DRAM: CMOS DRAM**
- **PCM: Phase Change Memory**
- **HRM: Horizontal Racetrack Memory**
- **VRM: Vertical Racetrack Memory**

	Bits / Cell	Cell Size (F**2)	Efficiency
Feature Size (nm)			
DRAM	1	6	0.65
PCM	4	6	0.65
HRM	10	4	0.65
VRM	100	4	0.65



Need to exploit more hardware concurrency per task

The decision on how this will be done is likely to play out by how the big machines in 2012 will exploit the hardware concurrency.

- **Threads appears as the clear leader at this time (applications are being threaded for existing and future machines now)**
- **Commercial is also moving in this direction.**
- **Dramatically changes the acceptable memory/perf relationship and allows for much higher performance per node. (a necessary but not sufficient condition)**
- **While this opens a new dimension…. There are new challenges to solve. Need to address all of these.**
    - **Having much more performance/chip must be balanced in terms of memory and I/O bandwidth.**
    - **Ability to efficiently scale threading to large numbers of threads is differentiator and enabler to exceptional power and cost efficiency.**
- **Need to focus on what type of threads and what type of hardware/software support is needed to get applications (commercial and HPC) to thread efficiently.**
- **Need to consider load imbalance, Amdahl, false cache line sharing, synchronization…**
- **Coarse grain threading is not easy but it is manageable.**
- **We can continue to look forward to additional "transparent" benefits from compilers and hardware speculative support.**

Cache and locality (distance is power)

- There is potential power to be saved by explicitly managing locality.
- But …. It can also easily go the other way. Best solution is to have the flexibility to do it either way.

	Relative Power in Dcache (detailed design study)			
	8-way	4-way	2-way	1-way
dcache	1.00	0.72	0.57	0.50
ddir	0.49	0.33	0.26	0.22
d set predict	0.17	0.08	0.04	0.00
derat	0.34	0.34	0.34	0.34
Total	1.99	1.47	1.21	1.06

Speculative data access

Baseline Power

Fundamental cache overhead

Latency improvement (possibly power)

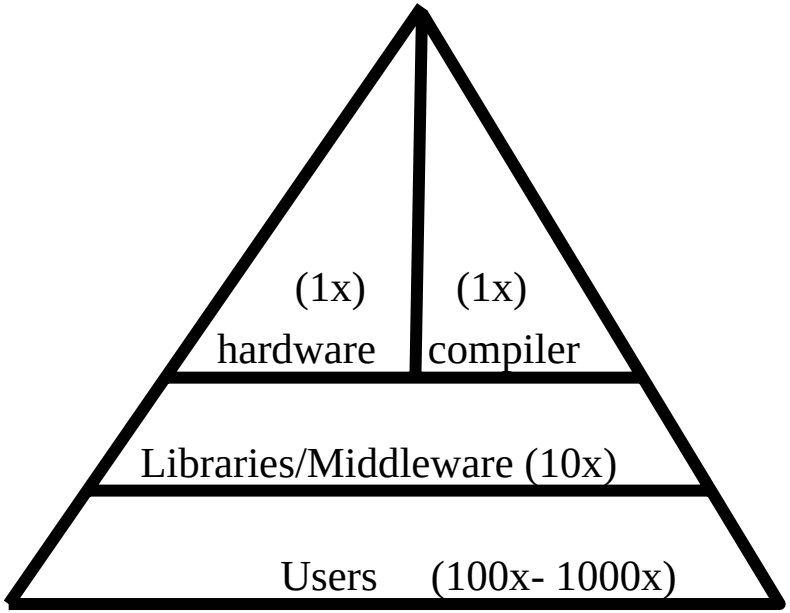
- In our Blue Gene systems that had this flexibility (L2) … little was actually gained by explicit control and this feature was rarely used.
  - If it is a frequently used data element cache replacement will tend to result in a cache hit.
  - If it is rarely used then there is little performance impact of having it in the cache.
  - Prefetching is a different story…. Clear value there.
  - Gather with memory remap can also have value. No really a memory versus cache issue

Reliability and impact on users

- The vast amount of silicon will make reliability a more difficult challenge.
- Multiple directions we can go here… it is up to the user community to decide

  Option 1) Leave it to system hardware and software to guarantee correctness. Not impossible, just expensive

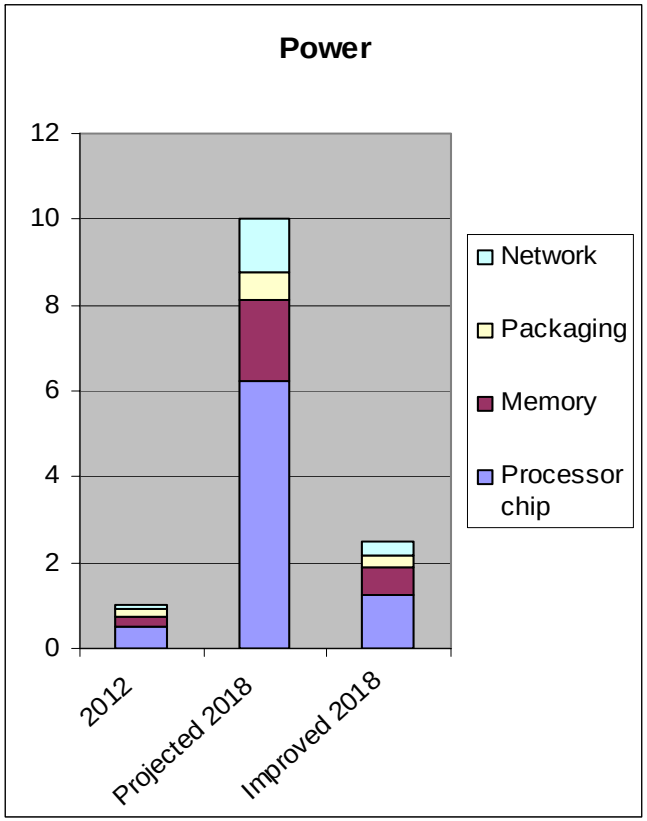
  Option 2) Leave it to the users to deal with potential hardware faults.

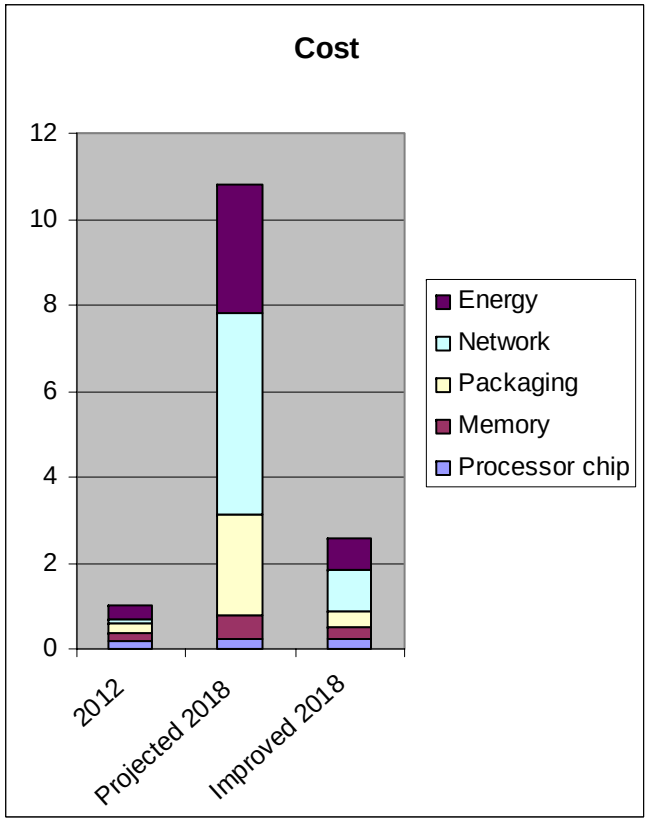


•Key to scalability is to keep it simple AND predictable.

•Keep reliability complexity away from the user as that is where the real cost is.

Memory, Bandwidth, and Network are significant contributors at Exascale
Bandwidth requirements drive up both Packaging and Network Costs





Optics cost drives the need for technology innovation as well as system network topology

- Reasonable Cost Target
  - Need to get to higher levels of integration to achieve better cost efficiency
  - Need to exploit new technologies to achieve even reasonable costs in some dimensions.
    - Optics : Required between nodes (Table 1) and may be useful on-node.
- Usability
  - Partnership is critical to making the right/optimal choices. Must have complete hardware/software solution.

Normalized to nearest neighbor communication, Other patterns will be different.

Network Topology	(3)Fully integrated silicon photonics. Risk factor HIGH ($M)	(2) Silicon photonics or Waveguides. Risk factor moderate ($M)	(1) VCSEL modules. Risk factor low ($M)	Comments
1-stage fat tree	120	360	840	not technically feasible
2-stage fat tree	240	720	1680	not likely technically feasible
3 stage fat tree	360	1080	2520	
4-D torus	60 (2MW)	135-180 (8-12MW)	420 (12MW)	(power, MW)
multidimensional switch	120	360	840	

Cost and power comparison of different network topologies for different levels of maturity and risk of optical technology. Costs are assumed to be for a machine in 2018 with a peak Exaflop and a byte/s-to-FLOP/s ratio of 1/10, and are estimated for off-node interconnects only.

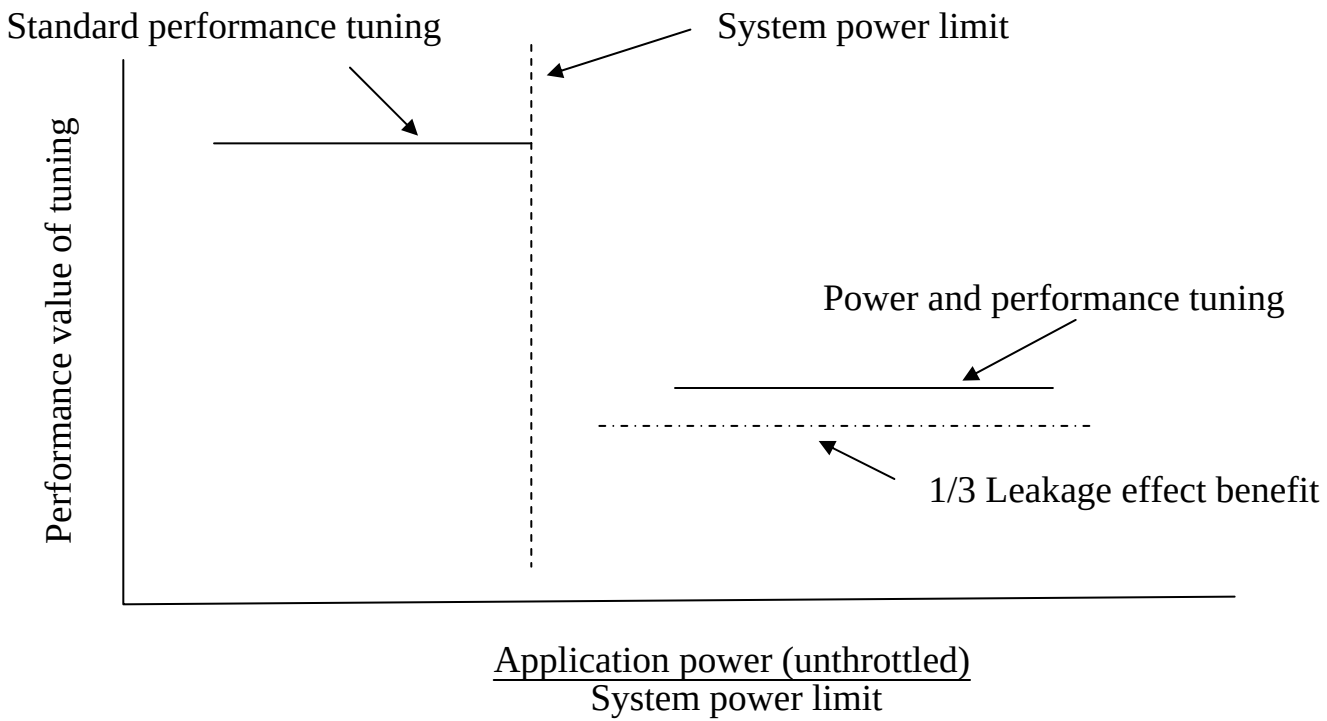
Energy constraints can change what "optimization" means.

- If sustained Perf/Watt is really the optimization point what will change…
  - Budget Watts not CPU hours?
  - With throttling does this mean that tuning code is a futile endeavor in the future?

Optimization technique	Perf optimized	Perf/Watt optimized
In Processor tuning. Eliminate pipeline stalls, unroll loops etc..	Yes	Little benefit . Leakage power amortized over more performance would be some positive impact
Strive for cache reuse at all level	Yes for latency and bandwidth	Yes for better Joules/Load
Pack data structures to get the most out of a cache line.	Yes for latency and bandwidth	Yes for better Joules/Load
Math functions (exp …) optimized using Newton methods	Yes for latency	Usually no… math is cheaper in terms of energy that memory loads
SIMD optimization	Yes	Yes, helps inside the processor
Prefetching within a node	Yes for latency	Probably little impact. Maybe negative. Possibly positive if performance significantly improved.
Software controlled power down/off of units	No	Yes
Indication of critical compute code path to system by user	No although compilers would like this. Compilers have a big challenge in this era.	Yes, prioritization of energy optimization is likely to help a great deal.
Overlapping of compute and communication	Yes for performance	a) No; Doesn't help as the network uses energy that will overlap with worst case compute. Better off separating them. b) Yes; Allows for slowing down compute and hiding it under communication.

Power limits will lead to a different type of tuning and degree of payback.

- Tuning of energy inefficient codes would follow traditional path
- Tuning of highly tuned code would be throttled back to hold to system limits.
  - Tuning focuses is on improving energy-performance
  - More performance will often mean more power so actual achieved higher performance will be harder to come by.
  - The higher the relative cost of power, the more motivation to push the system power limit to the left.



Summary

• Systems should provide a solid baseline with a workable incremental path to exceptional performance. (tools !!)

• Systems are very expensive and must be used simultaneously as production vehicles and as learning vehicles. Vendor – User partnership is critical.

• Threading models will likely be in production codes by 2012. We need to focus on the next layer… what type of threads.

• Breaking the memory per core or memory per thread constraint is a necessary but insufficient condition to a big cost efficiency jump. Breaking the JED

• Systems will need to be viable from a business perspective. (not just one or two instances)

• Optics is a very big issue for exascale (and we should expect no help from commodity on this)

•It is not all doom and gloom. HPC is growing. Systems will continue to advance as long as the results justify it.