Computing Requirements and Crystal Balls

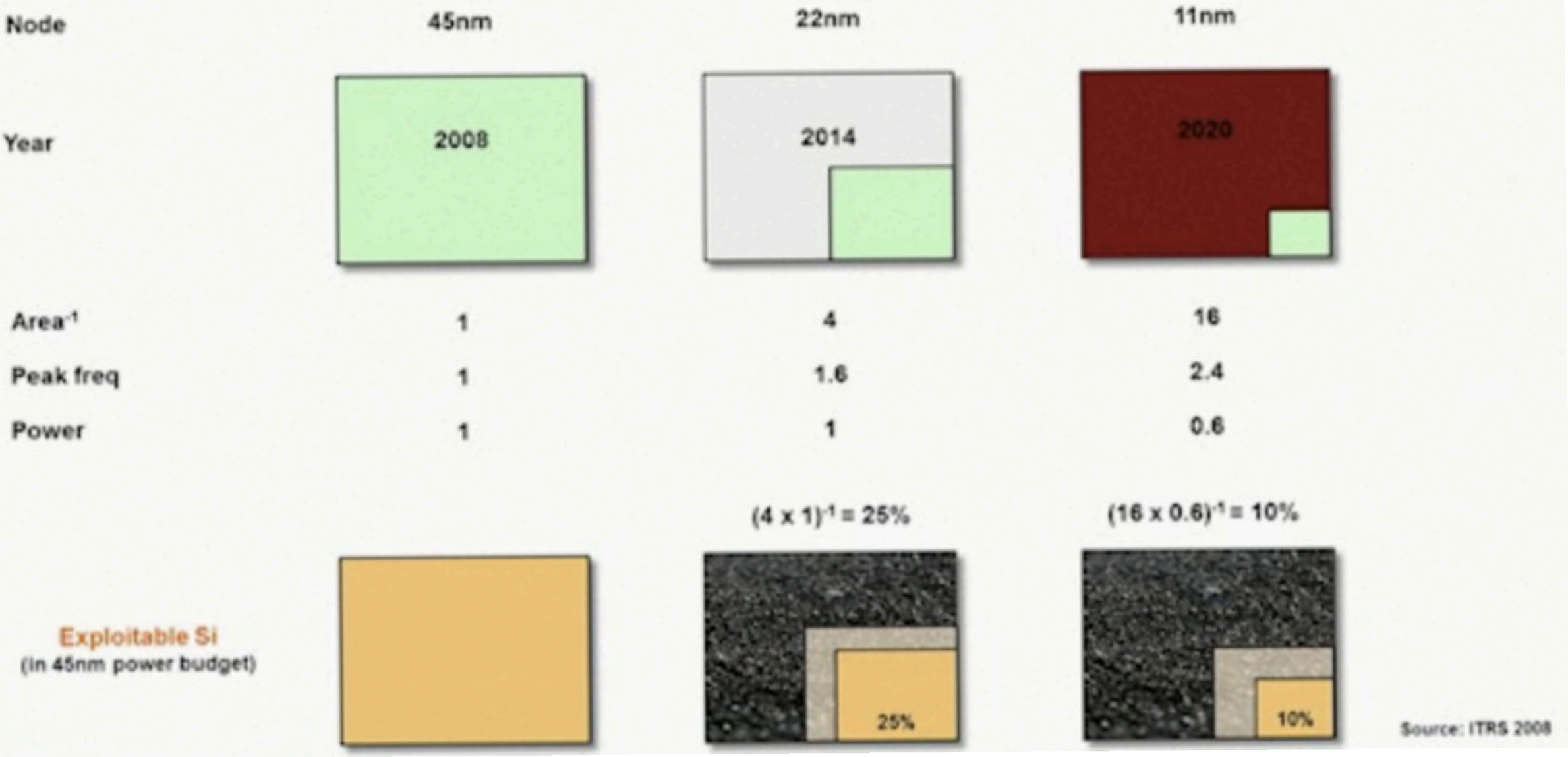
Davide Salomoni
INFN-CNAF

SuperB Computing R&D Workshop 2011
Ferrara, July 4-7, 2011

A World of “Laws”

- Moore’s law: CPU performance doubles every 18 months
  - Original version: transistor count doubles every two years.
- Kind of still working, but there’s considerably more to it
- Regardless of chip topology, multicore scaling is severely power limited
- Let’s all go many-cores, but two main points need to be taken into account:
  - Device energy efficiency is not scaling along with integration capacity.
  - What is the parallelism level of applications?

The Creation of Dark Silicon



- You **can** have more transistors, but you just can't **power** them all at the same time. (Jem Davies, ARM, at the AMD Fusion Developer Summit 2011)
- Conclusion: try and increase power management, but the road seems to point toward **heterogenous processing**: domain-specific processors to perform computing in the most efficient place.

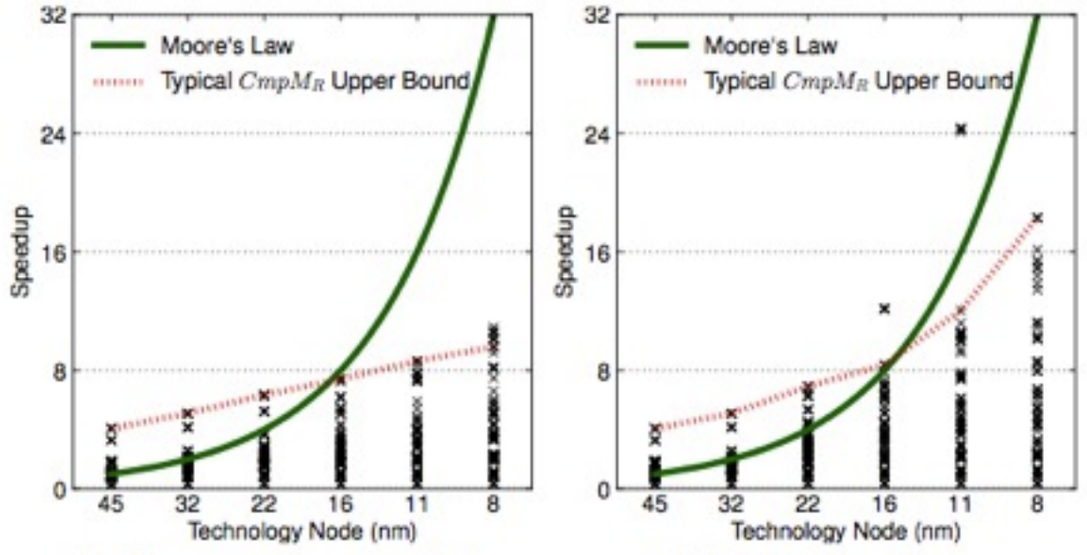
The Demise of Moore

- Cf. Esmaeilzadeh H., Blem E., St. Amant R., Sankaralingam K., Burger D., *Dark Silicon and the End of Multicore Scaling*, Proceedings of the 38th International Symposium on Computer Architecture (ISCA '11)
- Best-case average speedup of 7.9x between now and 2024 at 8nm
  - 16% annual performance gains for highly parallel workloads.
  - This is a 13x gap compared to Moore's law.
- A conservative scenario puts the speedup to a best-case average of 3.7x
  - Still for highly parallel workloads.
  - 22x gap compared to Moore's law.

Table 2: Scaling factors for ITRS and Conservative projections.

	Year	Tech Node (nm)	Frequency Scaling Factor (/45nm)	Vdd Scaling Factor (/45nm)	Capacitance Scaling Factor (/45nm)	Power Scaling Factor (/45nm)
ITRS	2010	45*	1.00	1.00	1.00	1.00
	2012	32*	1.09	0.93	0.7	0.66
	2015	22†	2.38	0.84	0.33	0.54
	2018	16†	3.21	0.75	0.21	0.38
	2021	11†	4.17	0.68	0.13	0.25
	2024	8†	3.85	0.62	0.08	0.12
	31% frequency increase and 35% power reduction per node					
Conservative	2008	45	1.00	1.00	1.00	1.00
	2010	32	1.10	0.93	0.75	0.71
	2012	22	1.19	0.88	0.56	0.52
	2014	16	1.25	0.86	0.42	0.39
	2016	11	1.30	0.84	0.32	0.29
	2018	8	1.34	0.84	0.24	0.22
	6% frequency increase and 23% power reduction per node					

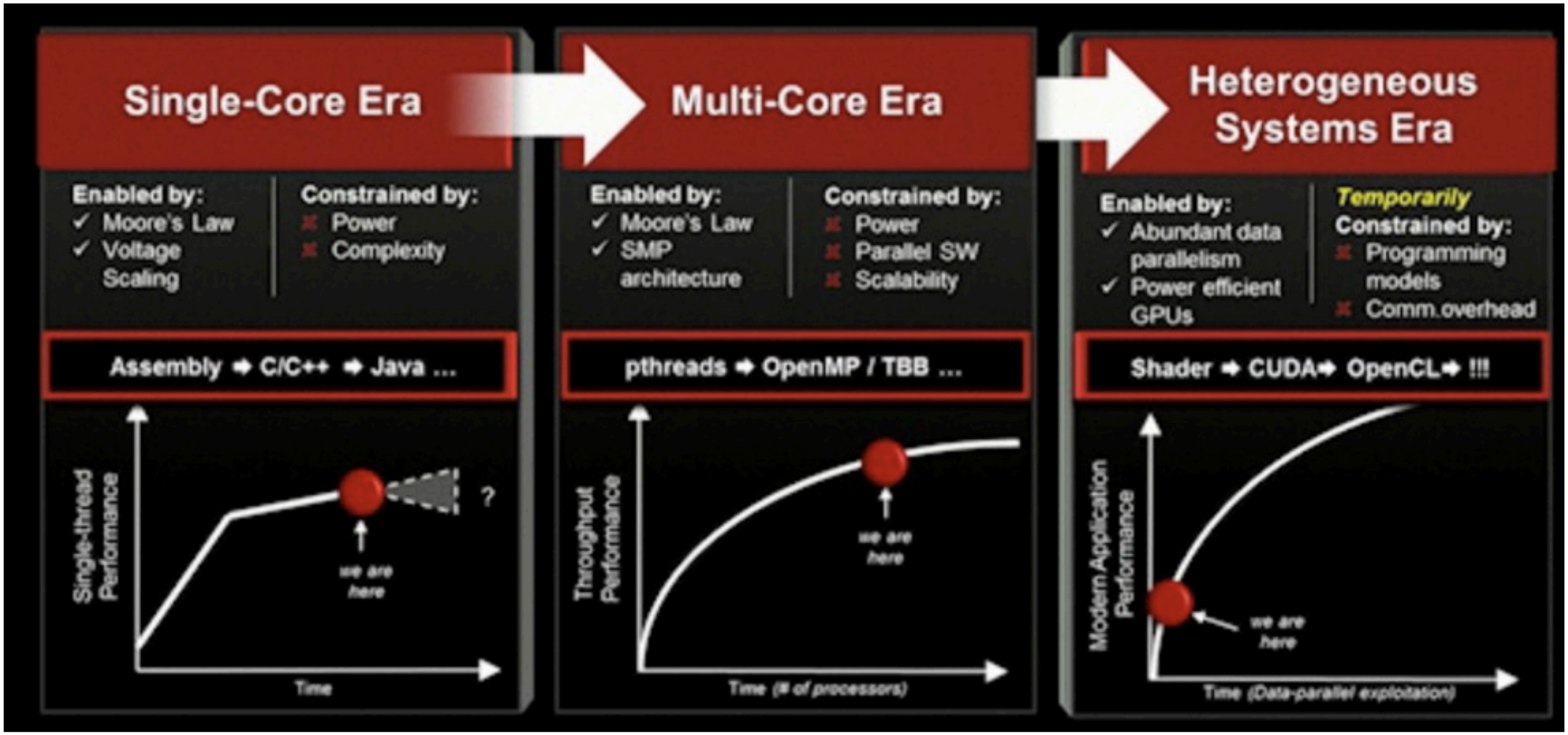
∗: Extended Planar Bulk Transistors, †:Multi-Gate Transistors



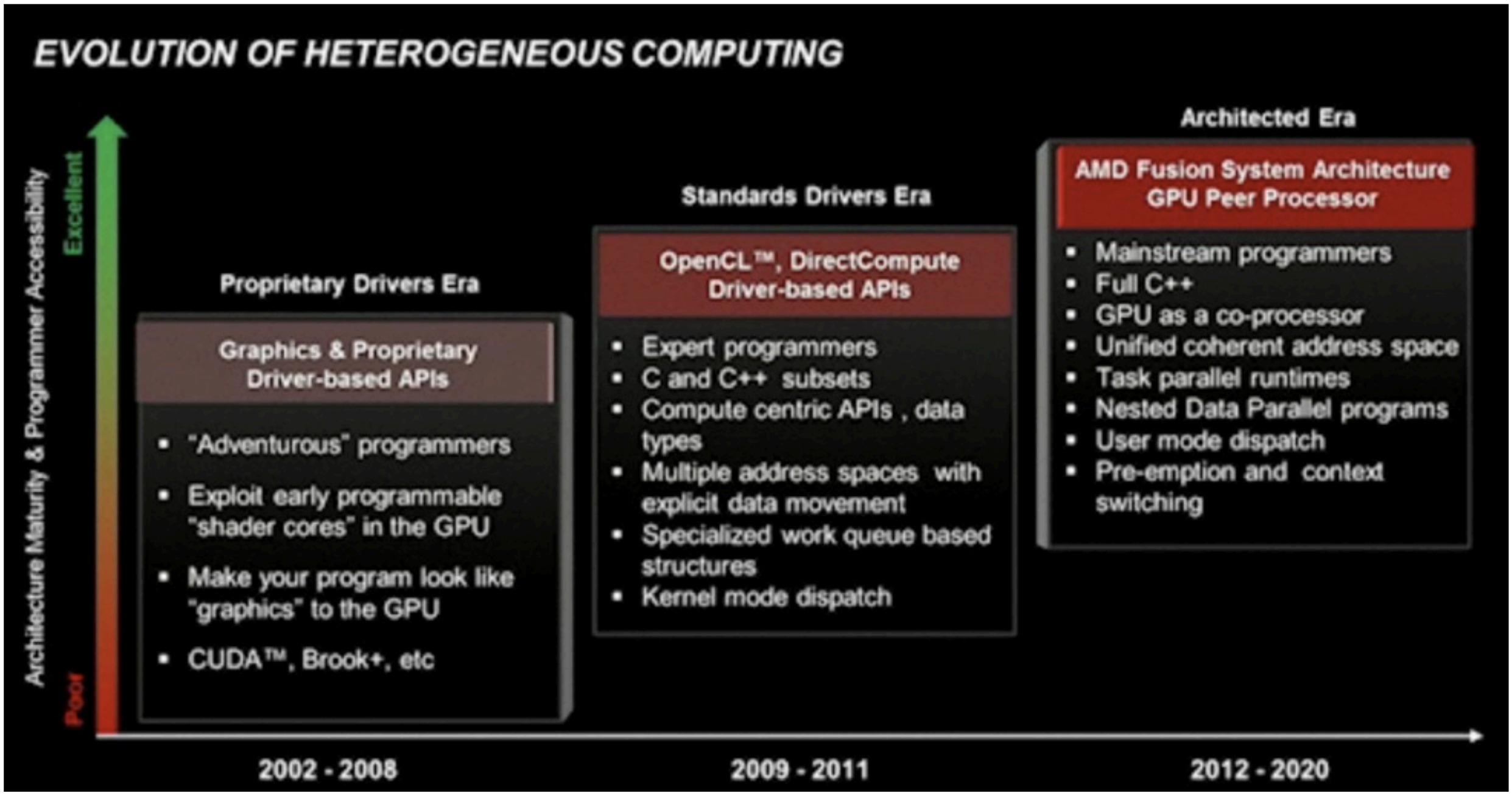
(a) Conservative Scaling (b) ITRS Scaling

Figure 9: Speedup across process technology nodes over all organizations and topologies with PARSEC benchmarks

Heterogeneous Computing



- Source: AMD Keynote, AFDS 2011



- Source: AMD Keynote, AFDS 2011

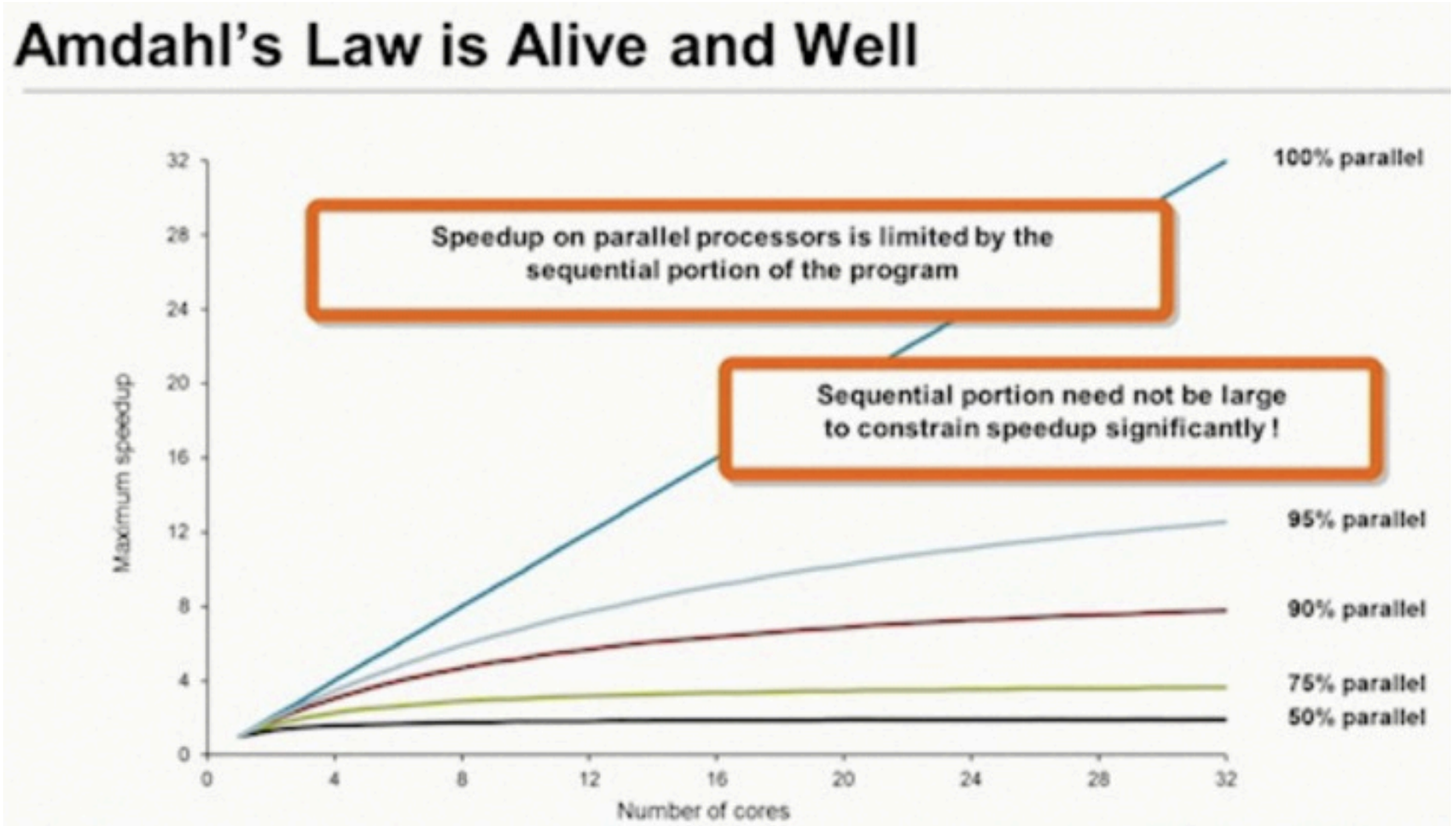
But GPUs...

- "1 TFLOP/s of [SP] performance in a single card"
  - Hence, it is easy and relatively cheap to build a "system" with hundreds of TFLOP/s.
- However:
  - GPUs are designed for high throughput and will typically only run well with very high numbers of simultaneous threads.
  - FLOP/s like all benchmarks tell only (if any) a part of the picture - one needs to take into account the "real system", i.e. FLOP/s vs. memory bandwidth vs. latency.
  - High parallelism is required, but what is "high"?

Amdahl's Law

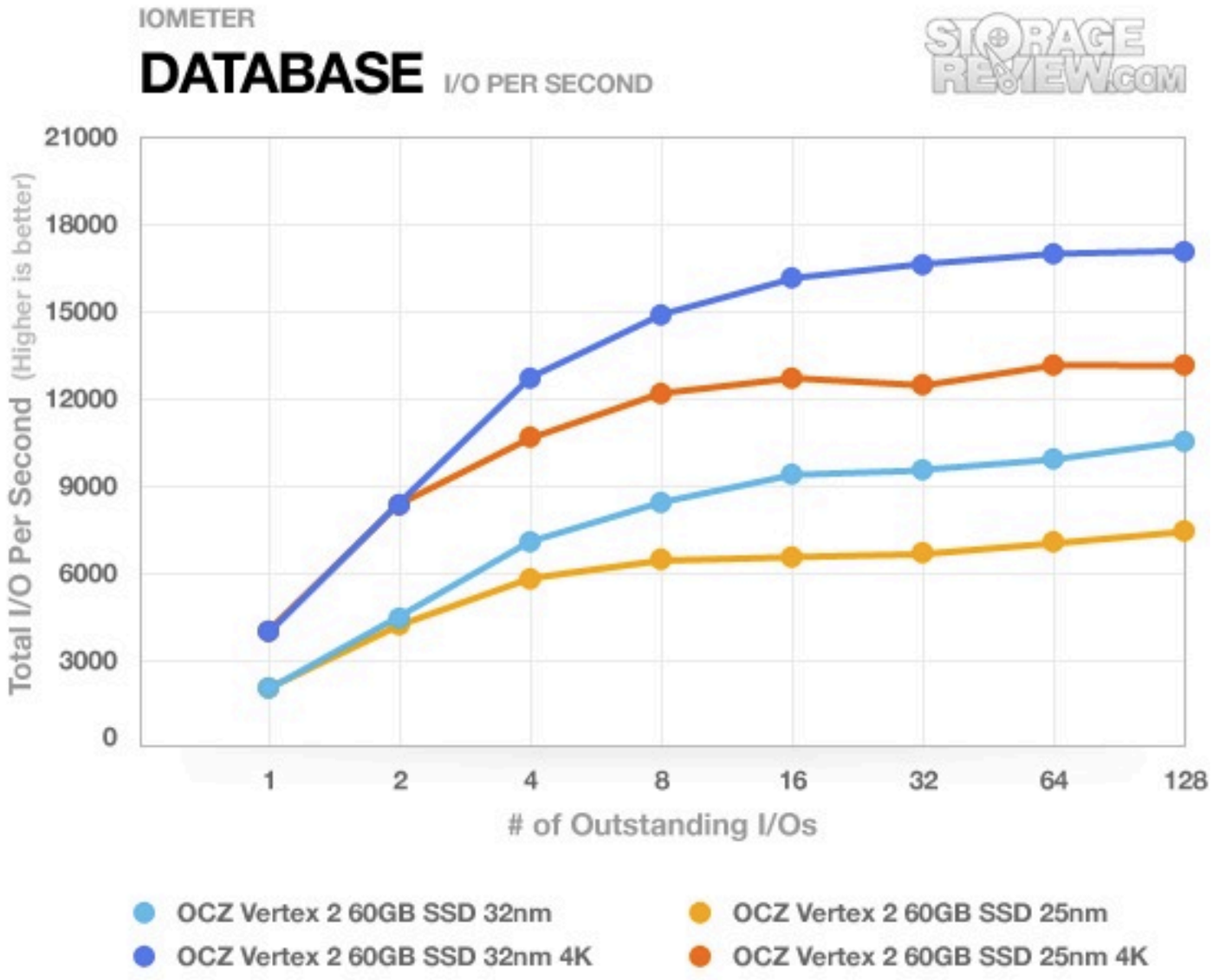
- The many-core model makes sense for highly parallel workloads
- If the problem size remains the same when parallelized, the max speedup S that can be achieved with N processors and a percentage P of parallel code is

$$S=\frac{1}{(1-P)+\frac{P}{N}}$$



Disks et al.

- Kryder's law ("Moore for storage"): disk storage density doubles every [year, or 18 months]
  - Good. However, even if the number of bytes on a disk that can be bought for unit cost follows Moore's law, **the speed of disk access does not**.
  - A possible solution may come from SSDs. However, there are still unresolved questions wrt for example **reliability and endurance**:
    - Need to increase capacity AND lower price per GB => use smaller processes
    - 32-34 nm NAND had 5000 write cycles
    - 25nm NAND down to 3000 write cycles; maybe not an issue for the consumer market, but other segments may think differently. This is "hidden" by some manufacturers with the decision to increase the SSD reserve capacity to cater for cells that will wear out - thereby reducing overall capacity for users.
    - **Performance does not necessarily increase with process reductions**.



Increasing Cores in "Traditional" Systems

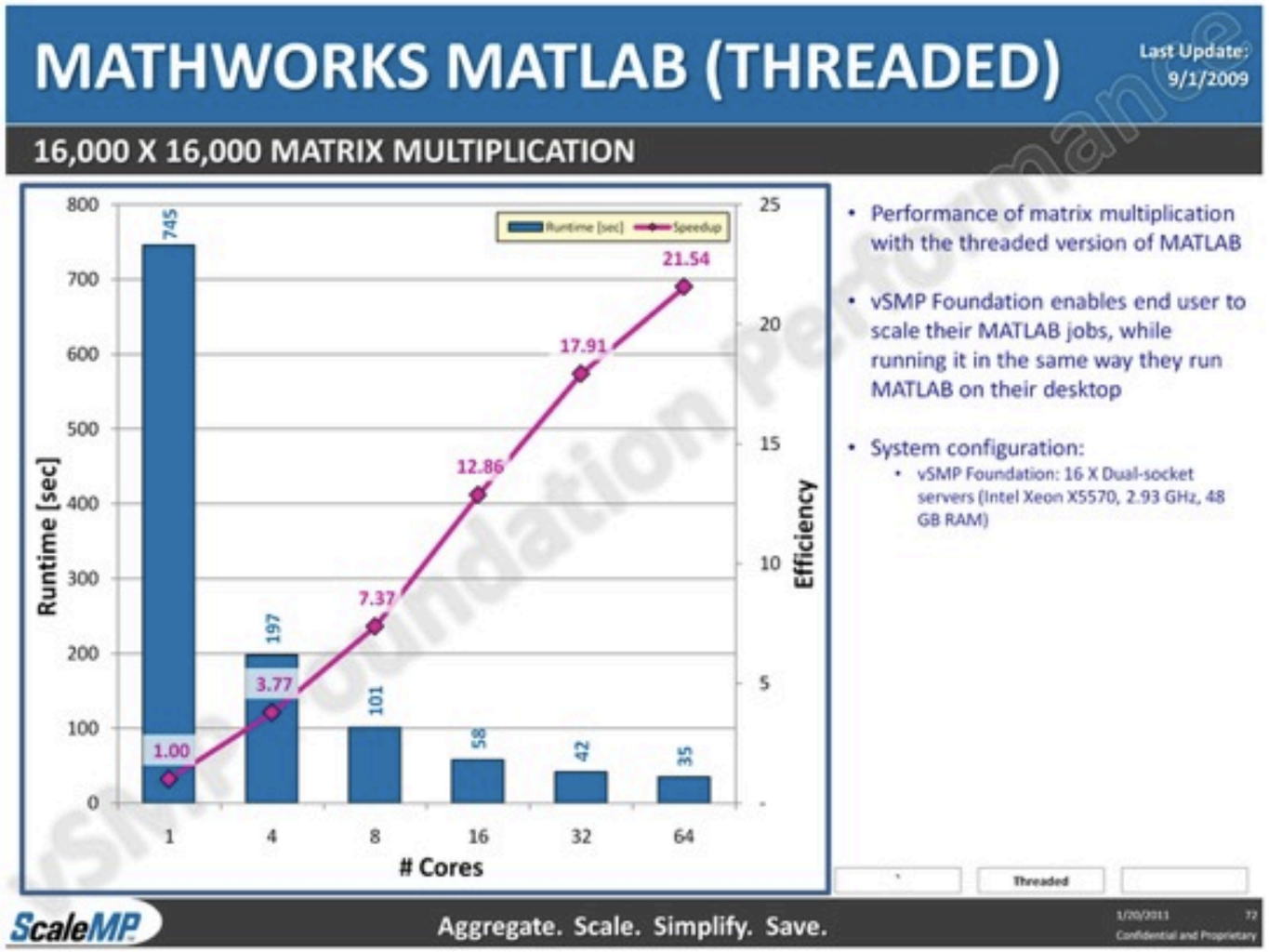
- How many disk drives? This is from a recent tender for CPU servers:

Amount of Physical Processing Cores	Number of Required Disk Spindle(s)
8	3
12	3
16	4
24	6
48	12

- The "whole nodes" debate: how are we going to manage shared-memory 48 cores (4-ways 12-cores CPUs) systems?
  - Applications' MP support is important, but consider also how normal, off-the-shelf services behave in these cases. (e.g. NFS, AFS caches with 100-200 GB RAM per system)

Partitioning Cores - or not?

- A solution to having “too many cores” per physical system may come from virtualization
  - This solves the “whole-and-too-many-cores-node” issue.
  - But still the virtual I/O area needs to significantly improve in performance for some applications.
- However, there may also be cases where a “whole-and-gigantic-node” is needed
  - This might be addressed with SMP systems created aggregating normal x86 systems
    - Cf. for example ScaleMP (commercial) or OpenSSI (open source).
  - Actually another form of virtualization
    - Which may exploit clustered filesystems



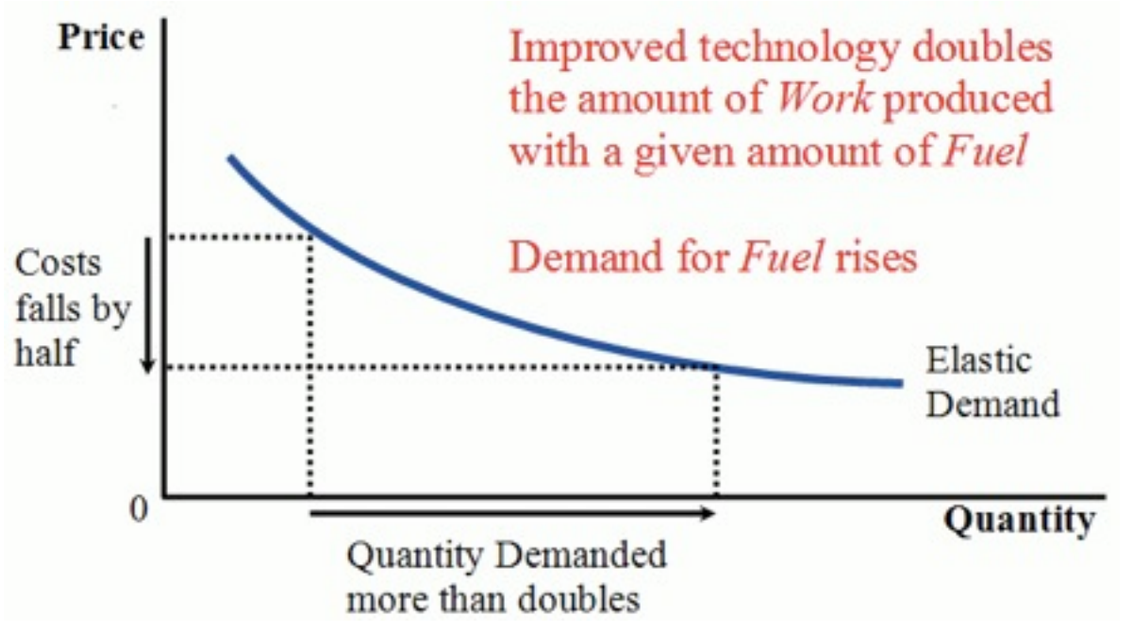
Networking

- Butter's law of photonics: the throughput of fiber optics doubles every nine months - and the cost of transmitting a bit over an optical network halves (it should, at least) every nine months
  - Where is the sweet spot of processing many-core data entirely over a local network? Diskless many-core CPU servers?
  - Can this be done on a (private) cloud? See table below from Armburst et al., *Above the Clouds: A Berkeley View of Cloud Computing*, EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-28.

	WAN bandwidth/mo.	**CPU hours (all cores)**	**disk storage**
Item in 2003	1 Mbps WAN link	2 GHz CPU, 2 GB DRAM	200 GB disk, 50 Mb/s transfer rate
Cost in 2003	$100/mo.	$2000	$200
$1 buys in 2003	1 GB	8 CPU hours	1 GB
Item in 2008	100 Mbps WAN link	2 GHz, 2 sockets, 4 cores/socket, 4 GB DRAM	1 TB disk, 115 MB/s sustained transfer
Cost in 2008	$3600/mo.	$1000	$100
$1 buys in 2008	2.7 GB	128 CPU hours	10 GB
cost/performance improvement	2.7x	16x	10x
Cost to rent $1 worth on AWS in 2008	**$0.27–$0.40** ($0.10–$0.15/GB × 3 GB)	**$2.56** (128× 2 VM's@$0.10 each)	**$1.20–$1.50** ($0.12–$0.15/GB-month × 10 GB)

And Finally...

- Wirth's (yes, Niklaus Wirth) law: software is getting slower more rapidly than hardware becomes faster.
  - E.g., Office 2007 performed the same task at half the speed on a year 2007 computer as compared to Office 2000 on a year 2000 computer.
- Software and hardware must evolve in parallel, *and be well matched*, to achieve greatness.
- Also, in economics there is something known as the Jevons paradox: there may be a rebound effect, so that bigger efficiencies lead to bigger (and not smaller) energy consumptions.
  - Let's not be *too* efficient (or successful) then! ☺



- Domain-specific processors should then be coupled with domain-specific development platforms.
- One could hope this will happen automatically, for example through intelligent compilers etc., but if this is not the case the effects of Moore's law (already challenged by many factors) will be further decreased.

Conclusions (1)

- **The future is apparently of many-core / heterogeneous computing**, but several concerns surround these concepts:
  - system balance: memory access, I/O and interconnect currently lack behind the increase in core count.
  - reliability: higher concurrency may well mean higher probability of failures.
  - programmability: which frameworks are we going to choose to (correctly!) use these many cores? Unified programming (e.g. à la Intel MIC Architecture) is theoretically very appealing.
  - efficiency and economics of distributed ("cloud, grid") computing: access to remote, possibly virtualized computing and storage (public, private) resources still needs to be properly modeled.
- **Moore's law itself seems challenged**
  - For example, it is expected that the TOP500 #1 system will reach 1 ExaFLOP/s by 2018, but today's #1 system (Tianhe-1A, China), would require more than 1.6 GW of power for that.

Conclusions (2)

- **For what regards SuperB computing**, I suppose that at this stage it may be difficult to fixate computing requirements with an uncertainty smaller than 2-4x. But whatever the estimates, the SuperB computing R&D framework should **carefully follow market trends** in the areas above and **establish widespread collaborations with industry and other partners** on this.
  - In particular, adapting ("optimizing") earlier software only shortly before data taking may not lead to efficient results.
- **And cost? This is crystal ballness at its best**.
  - Just as an example: in the past 5 years, Amazon EC2 progressively adopted newer CPUs for its customers. The CPU price decrease in this time period was in the order of 80 percent. However, the hourly rate for EC2 instances got a price reduction of about 15 percent only. (Vermeersch K., MD Thesis, Universiteit Antwerper, 2011) Many factors need to be taken into account here, including competition, global economics, overhead, and last but not least the cost of following trends and, learning, writing and adapting software for the new frameworks.