Will it be worthwhile to do science in the cloud?

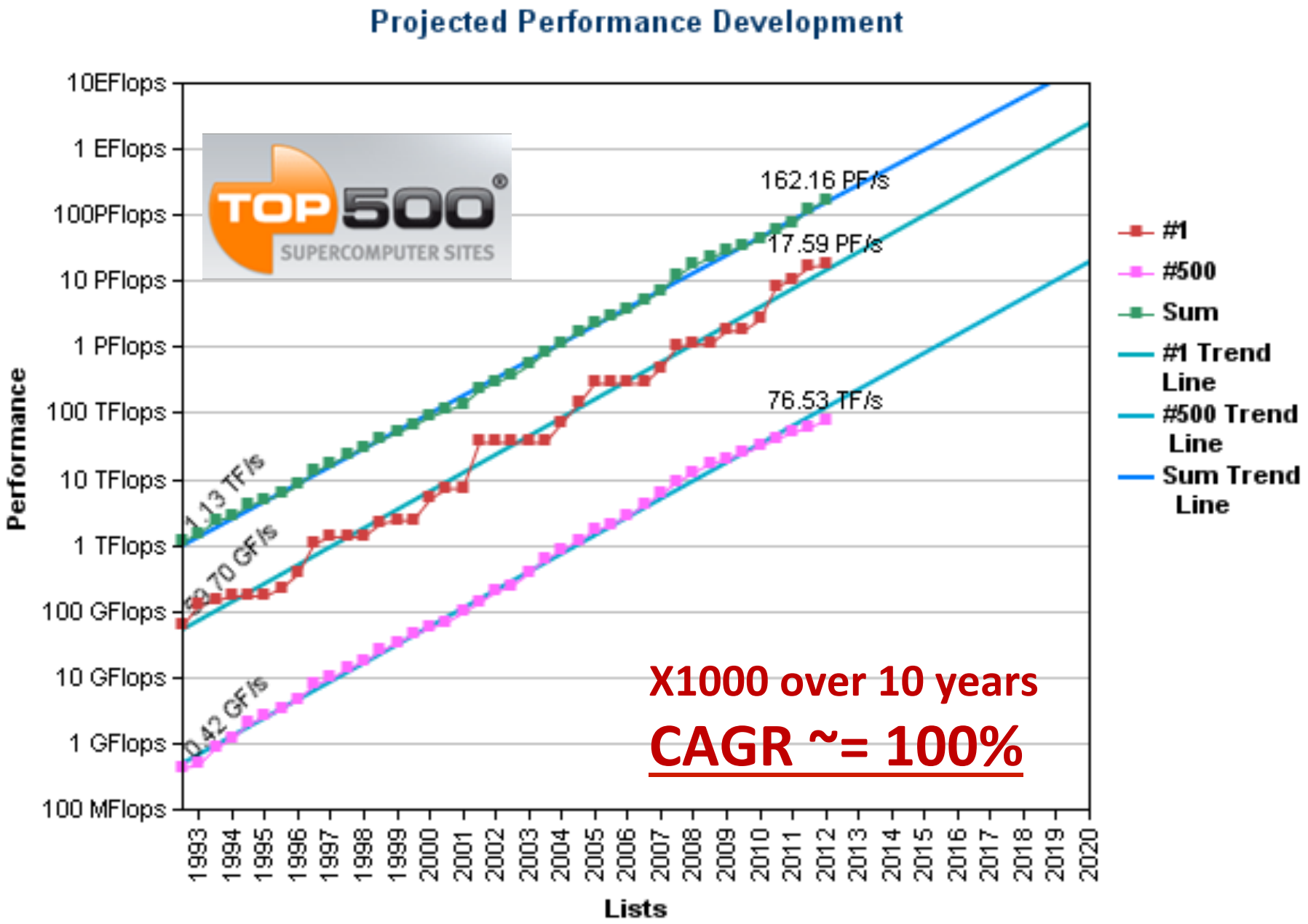
(Or, can we just get rid of those supercomputer centers and lease cycles and storage off Google, Amazon…)

Satoshi Matsuoka
Tokyo Institute of Technology

CCGrid2013 Panel

Science in the Cloud?

- Amazon HPC instance, Google Exacycle…
- "We don't need to invest in all that supercomputer R&D stuff; we invest into clouds, mobiles, big data, etc., and we will just leverage off those…"
- "Sure, developing apps are important, but as for the machines, sooner or later Google/Amazon/… will have enough resource in the cloud, so we will just use those…"


Projected Performance Development
TOP500
SUPERCOMPUTER SITES
10EFlops
1 EFlops
100PFlops
10 PFlops
1 PFlops
100 TFlops
10 TFlops
1 TFlops
100 GFlops
10 GFlops
1 GFlops
100 MFlops
Performance
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Lists
162.16 PF/s
17.59 PF/s
76.53 TF/s
1.13 TF/s
59.70 GF/s
0.42 GF/s
#1
#500
Sum
#1 Trend Line
#500 Trend Line
Sum Trend Line
X1000 over 10 years
CAGR ~= 100%

TSUBAME2.0 Nov. 1, 2010

"The Greenest Production Supercomputer in the World"

TOKYO TECH

TSUBAME2.0: A GPU-centric Green 2.4 Petaflops Supercomputer

Tsubame 2.0: "Tiny" footprint, very power efficient
- Floorspace less than 200m² (2,100 ft²)
- Top-class power efficient machine on the Green 500

TSUBAME 2.0
New Development

Chip
(CPU ,GPU)

CPU(Westmere EP)
76.8 GFLOPS
32nm

GPUs(Tesla M2050)
515 GFLOPS
3 GB
40nm

Compute Node
(2 CPUs,3 GPUs)

1.6 TFLOPS
55 GB/103 GB
>400GB/s Mem BW
80Gbps NW BW
~1KW max

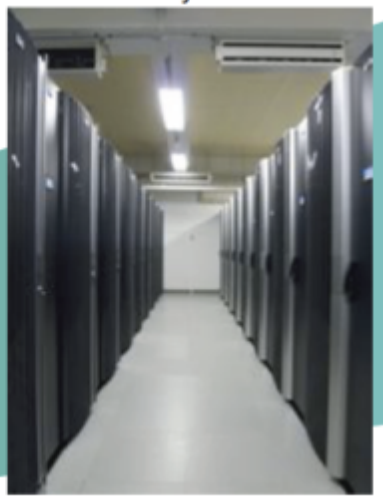
Node Chassis
(4 Compute Nodes)

6.7 TFLOPS
220 GB/412 GB
>1.6TB/s Mem BW

Rack
(8 Node Chassis)

53.6 TFLOPS
1.7 TB/3.2 TB
>12TB/s Mem BW
35KW Max

System
(42 Racks)
1408 GPU Compute Nodes,
34 Nehalem "Fat Memory" Nodes



2.4 PFLOPS
80 TB
>600TB/s Mem BW
220Tbps NW
Bisecion BW
1.4MW Max

Integrated by NEC Corporation

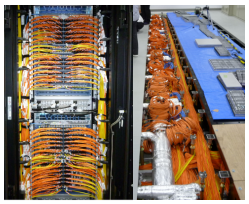

3500 Fiber Cables > 100Km
w/DFB Silicon Photonics
End-to-End 7.5GB/s, > 2us
Non-Blocking 220Tbps
Full Bisection Infiniband
40G single CMOS
Silicone Photonics

But what does "220Tbps" mean? Comparison with Entire Internet

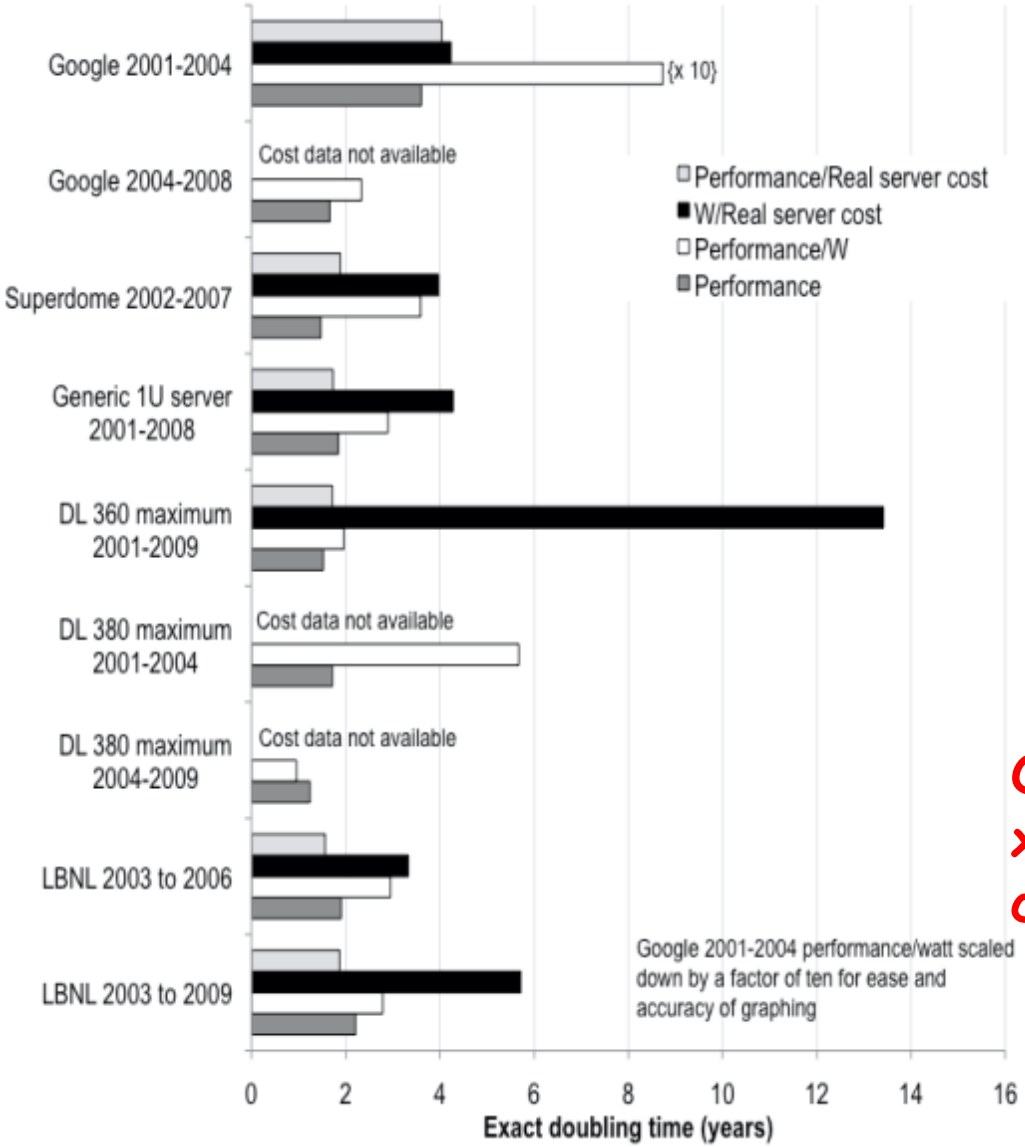
Global IP Traffic, 2011-2016 (Source Cicso)							
	2011	2012	2013	2014	2015	2016	CAGR 2011-2016
By Type (PB per Month / Average Bitrate in Tbps)							
Fixed Internet	23,288	32,990	40,587	50,888	64,349	81,347	28%
	71.9	101.8	125.3	157.1	198.6	251.1	
Managed IP	6,849	9,199	11,846	13,925	16,085	18,131	21%
	21.1	28.4	36.6	43.0	49.6	56.0	
Mobile data	597	1,252	2,379	4,215	6,896	10,804	78%
	1.8	3.9	7.3	13.0	21.3	33.3	
Total IP traffic	30,734	43,441	54,812	69,028	87,331	110,282	29%
	94.9	134.1	169.2	213.0	269.5	340.4	

TSUBAME2.0 Network has TWICE the capacity of the Global Internet, being used by 2.1 Billion users

Service rate proportional to bandwidth



Modern Server Speedups- x2 in 2 years

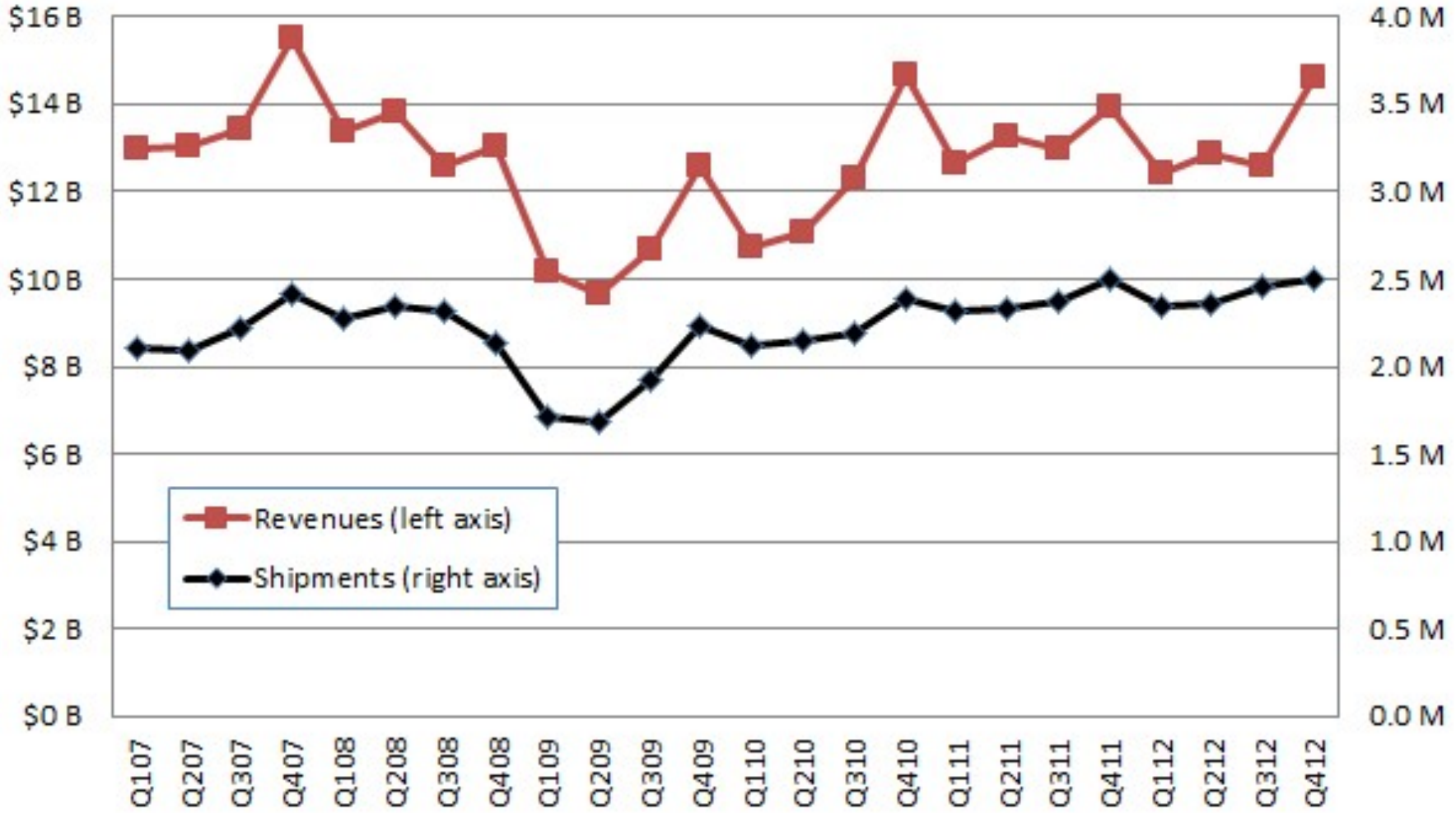


Source: **Assessing trends over time** in **performance**, **costs**, and **energy use** for **servers**, **Intel, 2009.**

"Performance per server and performance per thousand dollars of purchase cost double every two years or so, which tracks the typical doubling time for transistors on a chip predicted by the most recent incarnation of Moore's law"

Only x32 in 10 years -> x30 discrepancy c.f. x1000 in 10 years

Global Server Shipments are Flat – ~40% Capacity Growth Rate (~30% for non-HPC)



Service rate proportional to bandwidth

DoE Exascale Parameters
x1000 power efficiency in 10 years

System attributes	"2010"		"2015"		"2020"	
System peak	2 PetaFlops		100-200 PetaFlops		1 ExaFlop	
Power	Jaguar 6 MW	TSUBAME 1.3 MW	15 MW		20 MW	
System Memory	0.3PB	0.1PB	5 PB		32-64PB	
Node Perf	125GF	1.6TF	0.5TF	7TF	1TF	10TF
Node Mem BW	25GB/s	0.5TB/s	0.1TB/s	1TB/s	0.4TB/s	4TB/s
Node Concurrency	12	O(1000)	O(100)	O(1000)	O(1000)	O(10000
#Nodes	18,700	1442	50,000	5,000	1 million	100,000
Total Node Interconnect BW	1.5GB/s	8GB/s	20GB/s		200GB/s	
MTTI	O(days)		O(1 day)		O(1 day)	

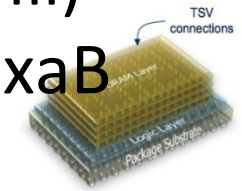
Billion Cores

Challenges of Exascale (FLOPS, Byte, …) (10^{18})!

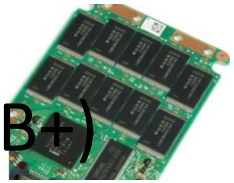
Various Physical Limitations Surface All-at-Once

- # CPU Cores: 1Bil Low Power — c.f. Total # of Smartphones sold globally = 400Mil
- # Nodes 100K~*x*M — c.f. The K Computer ~100K Google ~ 1 Mil

- Mem: *x*00PB~ExaB — c.f. Total mem all PCs (300Mil) shipped globally in 2011 ~ ExaB
  BTW $2^{64\sim}=1.8\times10^{19}=18$ExaB



- Storage : xExaB c.f. Google Storage 2 Exabytes (200Mil x 7GB+)



- All of this at 20MW (50GFlops/W), reliability (MTTI=days), ease of programming (billion cores?), cost… in 2020?!

What does this all mean?

- "Leveraging of mainframe technologies in HPC has been dead for some time."
- But leveraging non-HPC Cloud/Mobile sufficient for supercomputing for science?
- NO! They are already falling behind, and will be perpetually behind
  - CAGR of Clouds 30%, HPC 100%: all data supports it
  - Stagnation in network, storage, scaling, …
- Rather, HPC will be the technology driver for future IT, for Cloud/Mobile/Big Data to leverage