A lower bound to energy consumption of an exascale computer

Luděk Kučera
Charles University
Prague, Czech Republic

HPC'2014 Workshop, Cetraro, Italy, July 8, 2014


HOTAIR

Top5 (June 2014)

		Pflop/s*	MW	Cores
1. Tianhe-2	NUDT (China)	**33.9**	**17.8**	3,120,000
2. Titan XK7	Cray (USA)	**17.6**	**8.2**	560,640
3. Sequoia	IBM (USA)	**17.2**	**7.9**	1,572,864
4. K	Fujitsu (Japan)	**10.5**	**12.7**	705,024
5. Mira	IBM (USA)	**8.6**	**3.9**	786,432

* Linpack Benchmark

ExaScale Challenge

Build a system that performs 1 ExaFlop/s

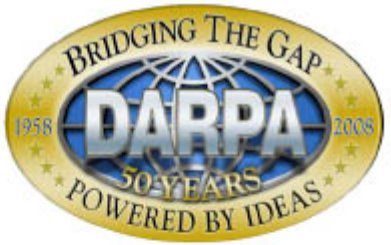
i.e., 10^{18} arithmetic operations per second
with double precision floating point numbers

i.e., 30 times more than Tianhe-2
and more than 50 times faster than Titan Cray

When? Soon – in 2015 ! (???)

ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems

Peter Kogge, Editor & Study Lead
Keren Bergman
Shekhar Borkar
Dan Campbell
William Carlson
William Dally
Monty Denneau
Paul Franzon
William Harrod
Kerry Hill
Jon Hiller
Sherman Karp
Stephen Keckler
Dean Klein
Robert Lucas
Mark Richards
Al Scarpelli
Steven Scott
Allan Snavely
Thomas Sterling
R. Stanley Williams
Katherine Yelick



2.2.1 Data Center System

For this study, an **exa-sized data center system of 2015** is one that roughly corresponds to a typical notion of a supercomputer center today - a large machine room of several thousand square feet and multiple megawatts of power consumption. This is the class system that would fall in the same footprint **as the Petascale systems of 2010, except with 1,000x the capability**. Because of the diffculty of achieving such physical constraints, the study was permitted to assume some growth, perhaps a factor of 2X, to something with a maximum limit of 500 racks and **20 MW for the computational part of the 2015 system.**

ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems

Peter Kogge, Editor & Study Lead
Keren Bergman
Shekhar Borkar
Dan Campbell
William Carlson
William Dally
Monty Denneau
Paul Franzon
William Harrod
Kerry Hill
Jon Hiller
Sherman Karp
Stephen Keckler
Dean Klein
Robert Lucas
Mark Richards
Al Scarpelli
Steven Scott
Allan Snavely
Thomas Sterling
R. Stanley Williams
Katherine Yelick

September 28, 2008


Ferrari


EL-716 WV


FIAT

Top 500 computers **together**

0.250 ExaFlop/s
only 32 of them have more than 1 PetaFlop/s

Only less than one half ot Top500 computers
report their power, but even those need more than 600 MW

How much energy we would need for an ExaFlop/s computer

Tianhe-2	33.8 PFlop/s	17.8 MW	i.e. 1.90 GFlop/J
Titan XK7	17.6 PFlop/s	8.2 MW	i.e. 2.14 GFlop/J
Sequoia	17.2 PFlop/s	7.9 MW	i.e. 2.18 GFlop/J

Assuming 2 Gflop/J, for 1 ExaFlop/s

we would need 500 MW

25 times more than the DARPA requirement

10 times more than many authors consider as feasible

Do we really need ExaScale?

Who builds 20 MW Exascale System (DARPA)

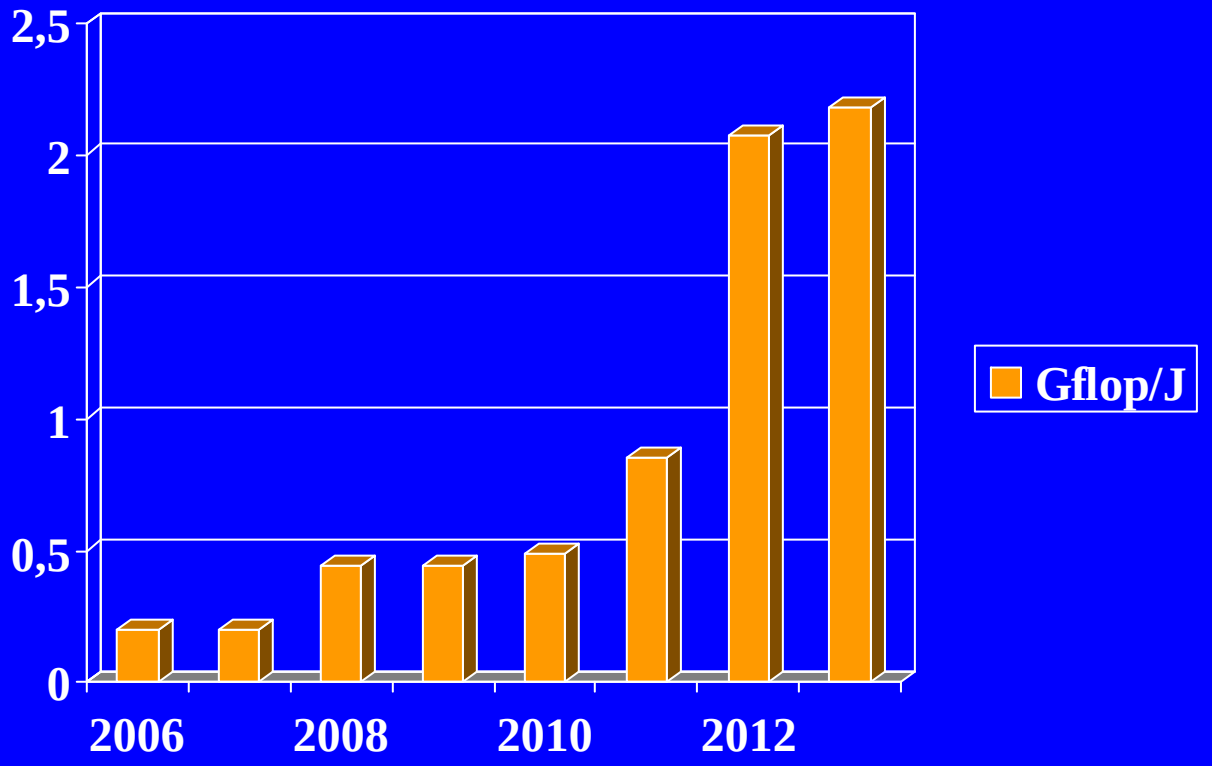
Knows building 20 kW 1.0 PetaFlop/s System
(it would rank 33rd in Top500/Nov2013)

and 0.125 PetaFlop/s System
as power hungry as a wash-machine (2.5 kW)
(it would rank 464th in Top500/Nov2013)

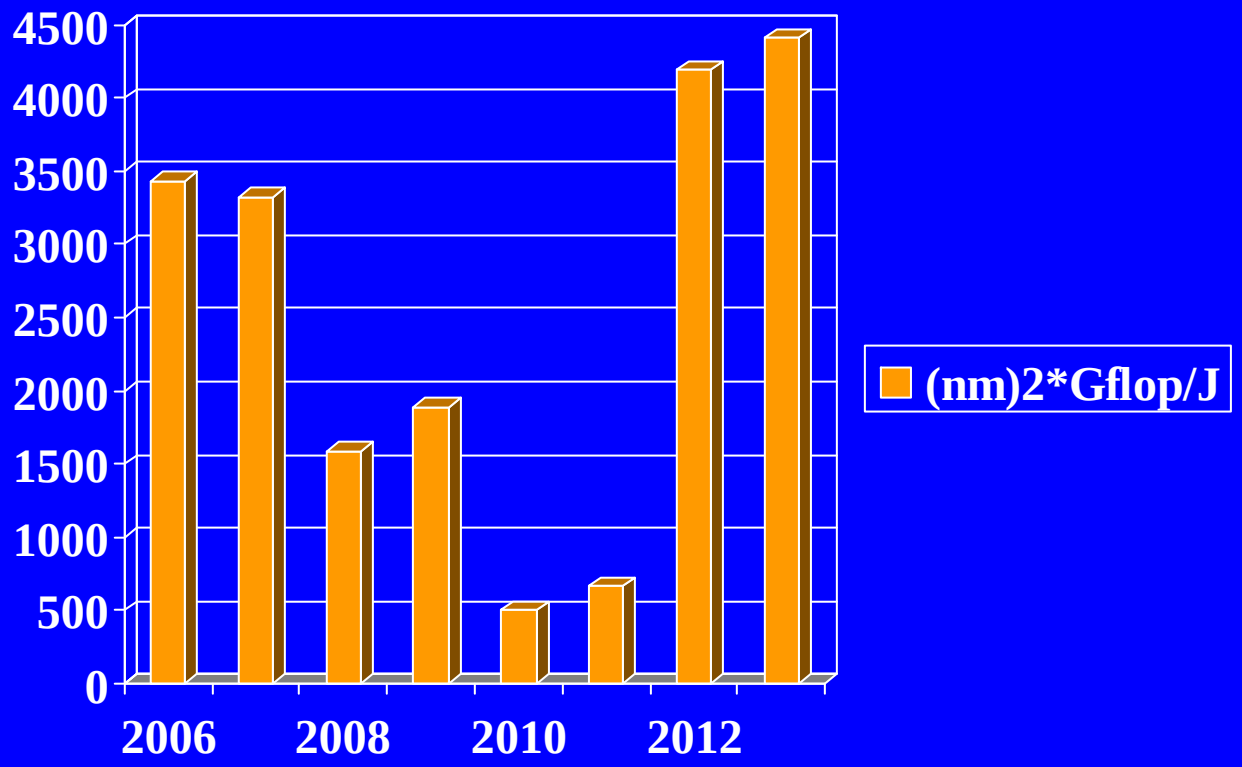
Evolution of Gflop/J

2005 BlueGene/L	0.2 GFlop/J	130nm
2006 BlueGene/L	0.2 GFlop/J	90nm
2007 BlueGene/L	0.2 GFlop/J	90nm
2008 IBM Roadrunner	0.44 GFlop/J	65nm
2009 IBM Roadrunner	0.44 GFlop/J	65nm
2010 Nebulae	0.49 GFlop/J	32nm
2011 Tsubame	0.85 GFlop/J	28nm
2012 Sequoia	2.18 GFlop/J	45nm
2013 Sequoia	2.18 GFlop/J	45nm

Evolution of Gflop/J



MFlop/J * (technology)^2



How much energy do we need for 10^{18} multiplications?

Forget CPU's, GPU's, Xeon's, Kepler's, etc.

Forget buses, caches, memory, interconnect

Forget static power requirements

Assume just a standard CMOS multiplier units and their dynamic energy needs

How much power do we need for 10^{18} multiplications/sec?

A) < 10MW

B) 10-50 MW

C) 50 –100 MW

D) > 100MW

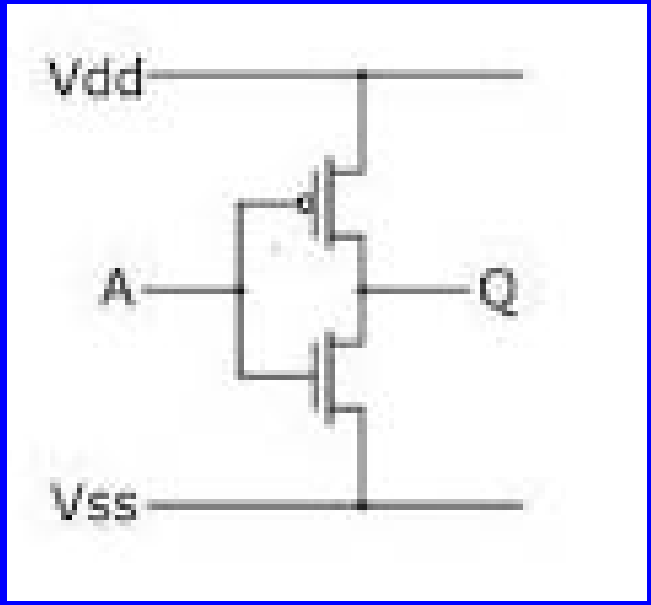
How much energy do we need for 10^{18} multiplications?

How many bit changes are necessary
for one multiplication

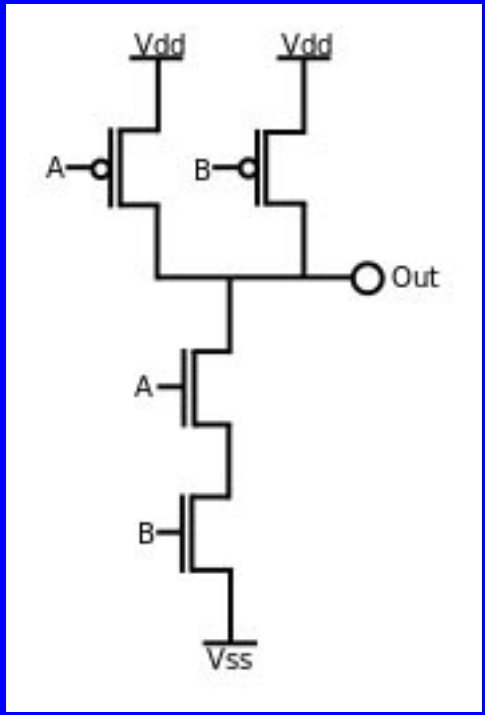
multiplied by the energy for one bit change

multiplied by 10^{18}

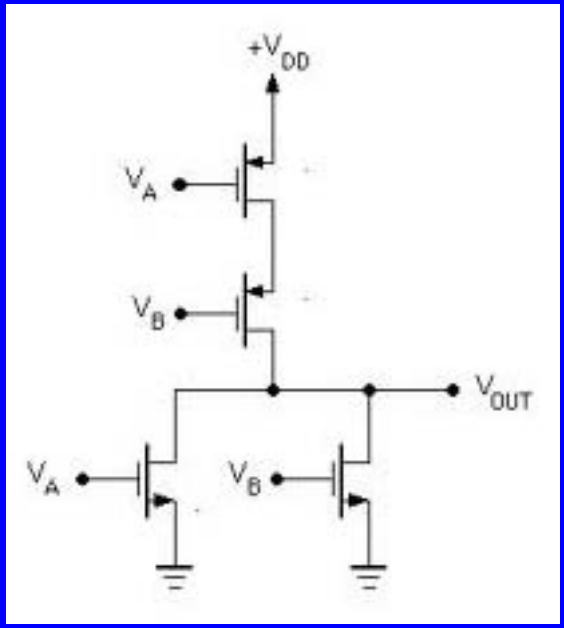
CMOS NOT gate (invertor)



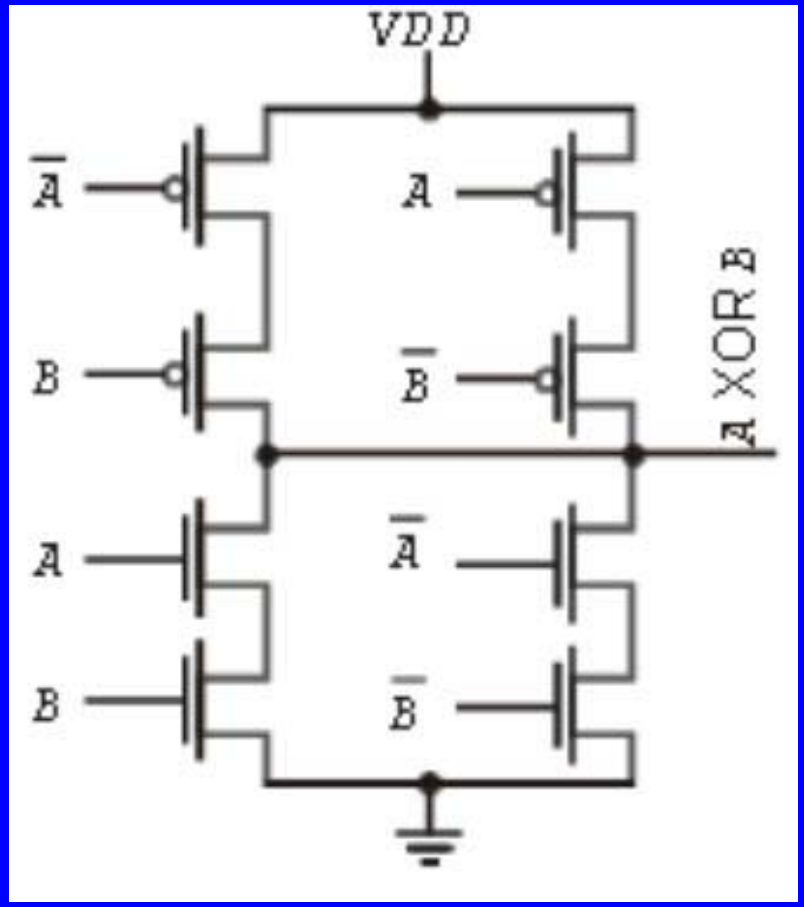
CMOS NAND gate



CMOS NOR gate



CMOS XOR gate



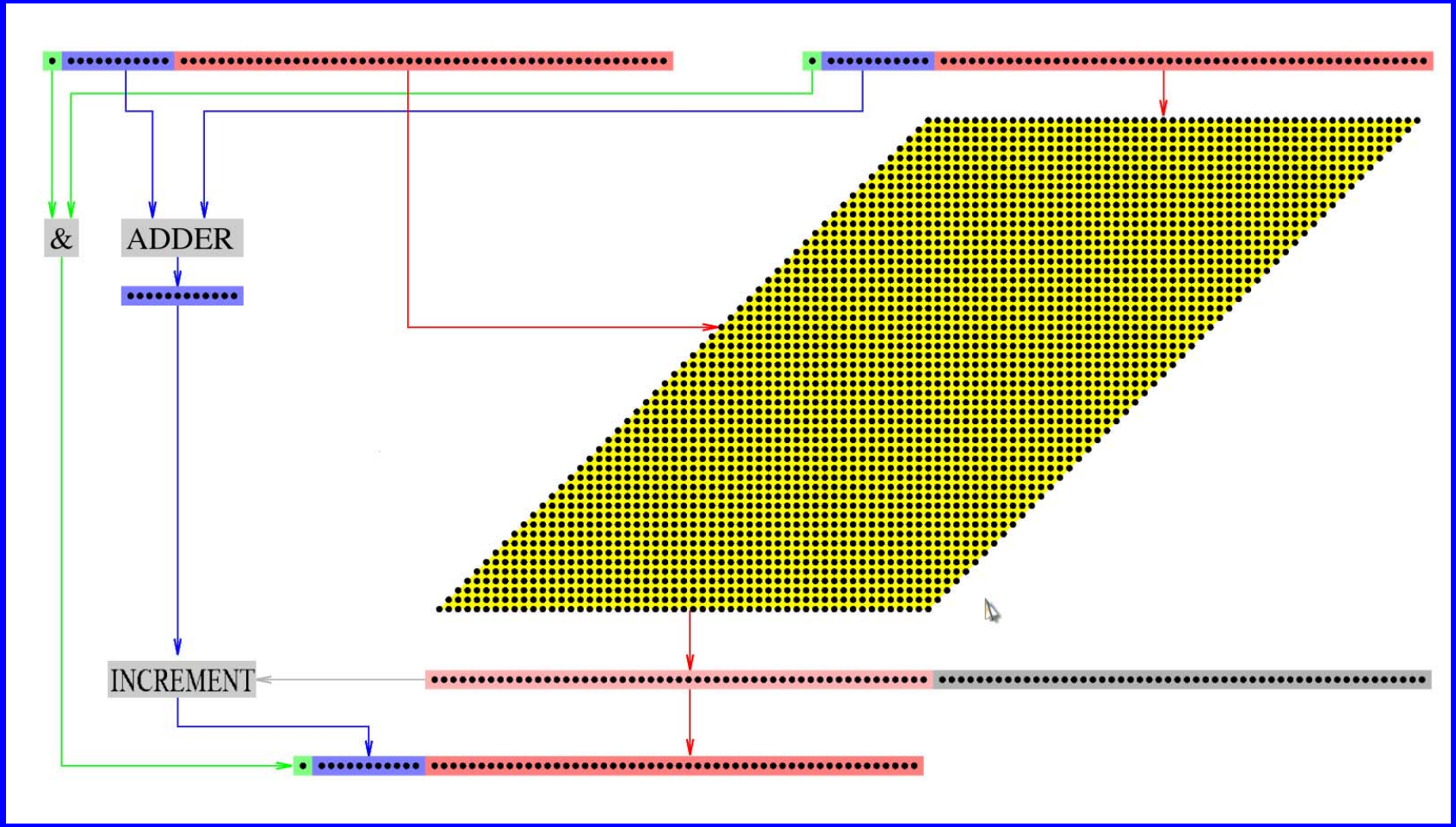
Double precision floating point number IEEE 754

1 sign bit

11 exponent bits

53 significant bits (52 explicitly stored)

Double precision floating point multiplier



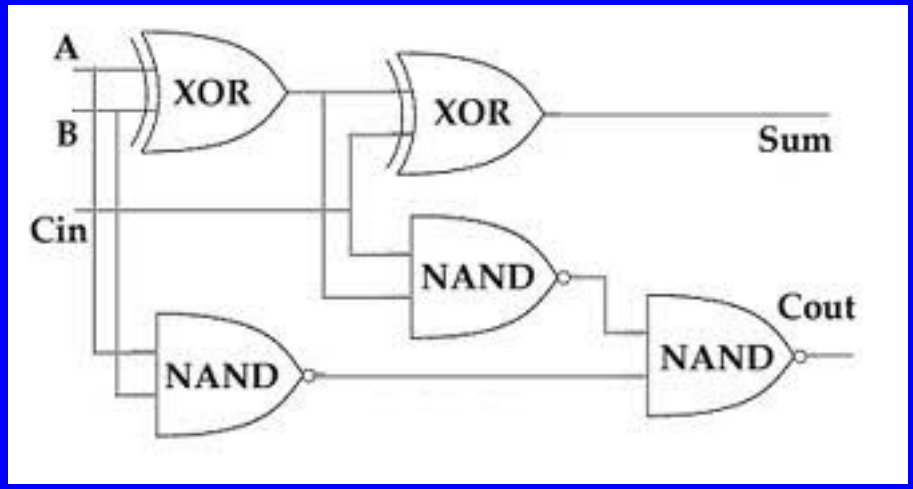
Wallace tree

A full adder is used as a compressor that transforms 3 items of the multiplication table into 2 items.

FA – 3 inputs (order k), 2 inputs (order k and order k+1)

To compress 53x53 = 2809 items of the multiplication matrix into 106 bits of the product we need more than 2700 full adders

Full adder



Bit changes per 1 multiplication (IEEE 754)

A particular (not very optimized) implementation of a IEEE 754 double precision floating point multiplier (using Wallace trees)

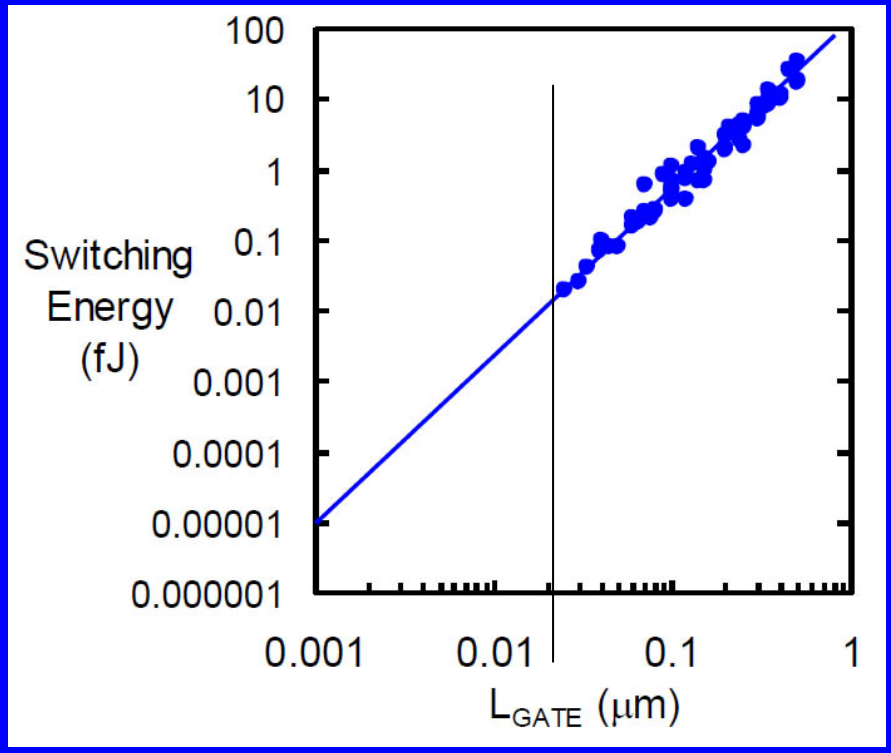
Randomly generated double precision numbers

Approximately **6000** bit changes / multiplication

NAND 2200 changes
NOR 1000 changes
XOR 2700 changes
NOT 100 changes

Current CMOS device scaling

Dmitri Nikonov, Intel Corp. (2013),
Course on Beyond CMOS Computing, https://nanohub.org/resources/18347.



Feasible energy – 1 fJ (femtoJoule = 10^{15} J)

Power estimation for 10^{18} IEEE 754 multiplications/sec

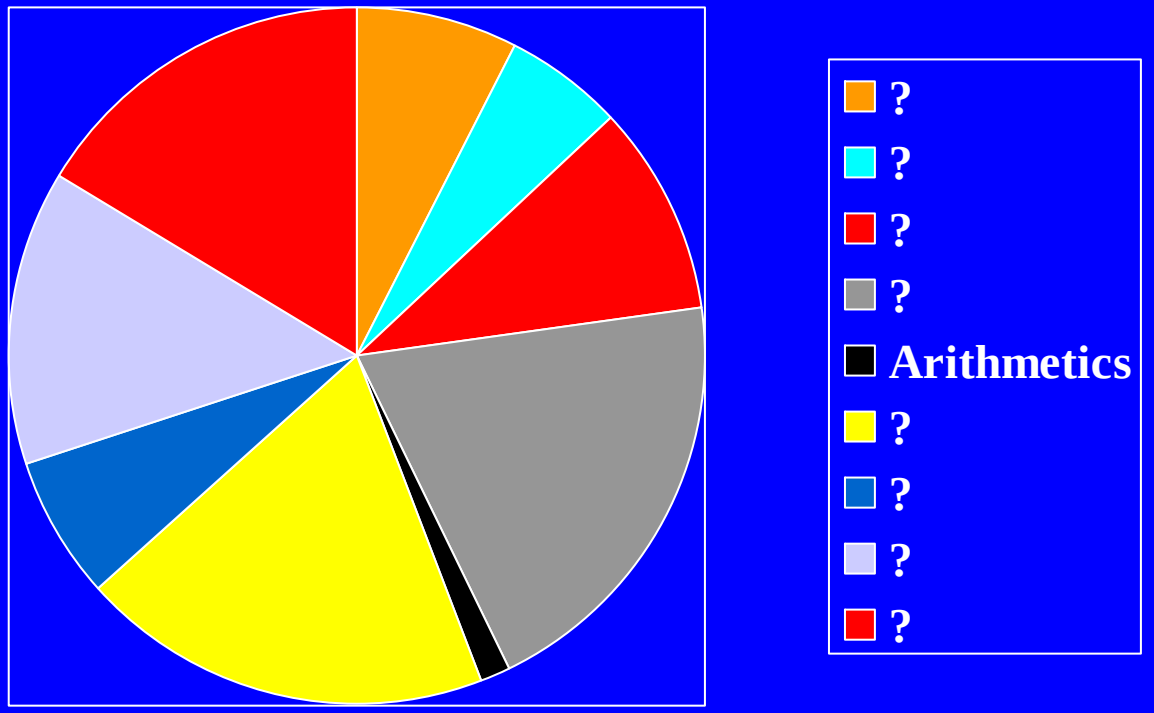
10^{18} x 6000 x 1 fJ / sec

$6 \times 10^{21} \times 10^{-15}$ J/sec

6 MW


WANTED
for eating
494 MW
Big
reward

Needed: Power consumption



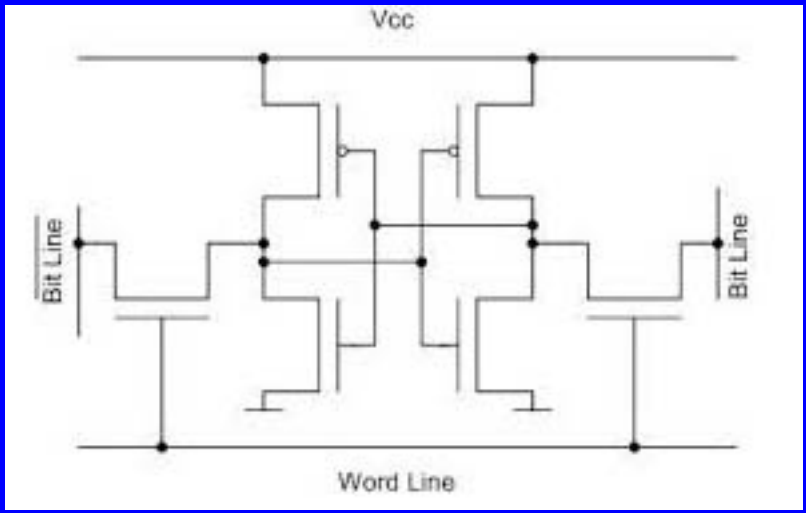
64 bit number storing

Worst: 64 memory cells change their state

Average: ~ 32 memory cells change their state

Compare to 6000 bit changes (on average)
for multiplication

CMOS static memory cell (6 gates)



Interconnect

A communication pattern is strongly dependent on a problem being solved

different levels of communication:

- within a processing unit (e.g., a multiplier)
- within a core (e.g., ALU – cache)
- among cores within a single chip
- within a board or a rack
- long distance communication

Conclusions

The analysis suggest that we don't need to go beyond CMOS to build an exascale system

The analysis suggests that we should re-think the architecture of computing chips and interconnect fabrics
To approach 50 GFlop/J from the present 2 Gflop/J

Thank you for your attention

and Lucio Grandinetti
for organizing HPC workshops