

Superconducting **COMPUTING** :

An energy-efficient quantum-based technology for supercomputers

Pascal FEBVRE

IMEP-LAHC - CNRS UMR 5130
Université Savoie Mont Blanc
France

Pascal.Febvre@univ-smb.fr

Acknowledgements : Akira Fujimaki, Deep Gupta, Juergen Kunert, Coenrad Fourie, Ziad Melhem, Thomas Ortlepp, Alain Ravex, Hannes Toepfer, Georges Waysand, Ugur Yilmaz, Nobuyuki Yoshikawa

data centres (Google, Facebook,...)

Google servers - The Dalles - Oregon



Copyright Google

Energy consumption

In 2010, routers and servers consumed :

- between 1.1 % and 1.5 % of the total energy production worldwide
- between 1.7 % and 2.2 % in the United States of America



Source : Google



Need of computers with higher performance

The demand of high performance computers and servers will continue to grow. We need :

- better predictive models for climate change and weather forecast ;
- to better understand the formation of the early Universe ;
- to understand subatomic physics ;
- to model cells, for genetics, biotechnologies ;
- to simulate brain functions, ...

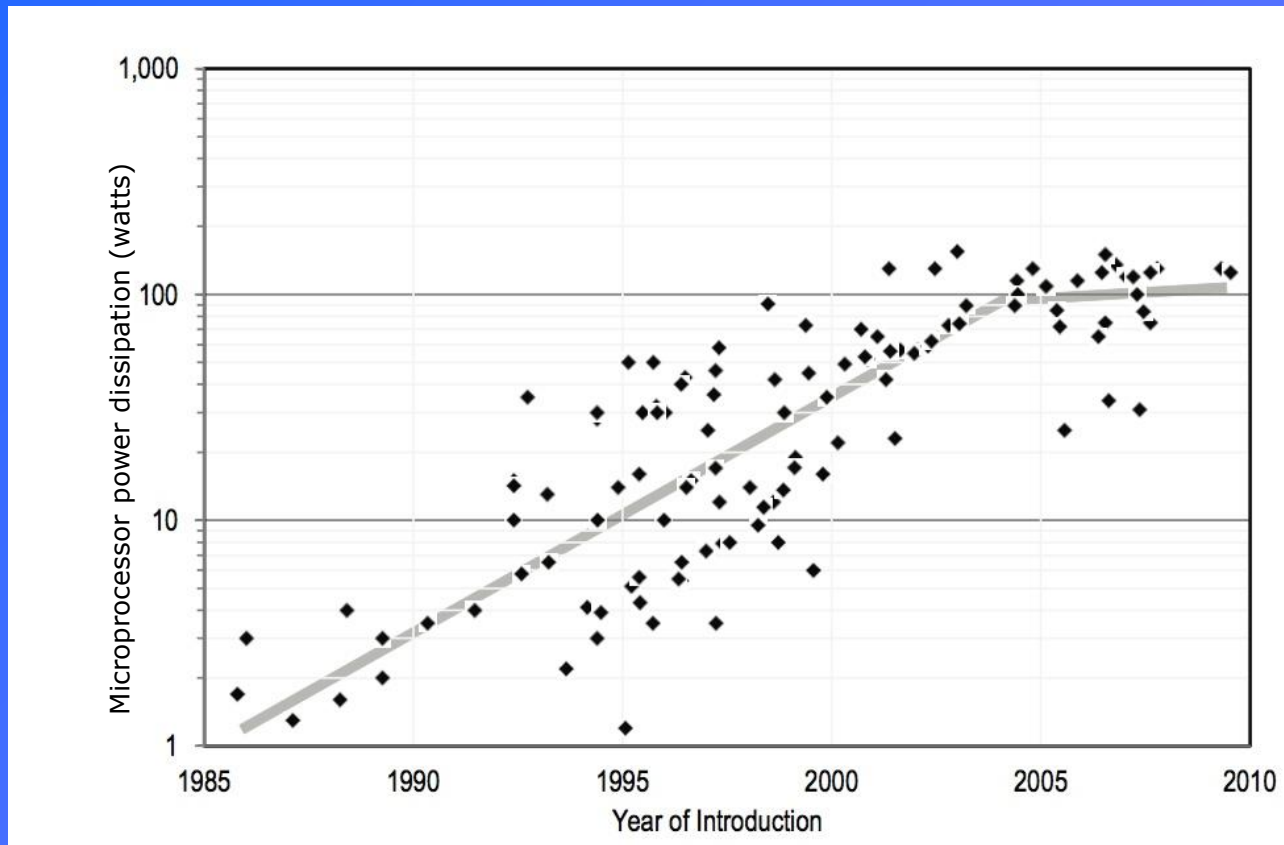


But the required power cannot follow the same pace

Dennard scaling law

An equivalent reduction of the power consumption per device is achieved, to keep constant the power dissipated by the chip.

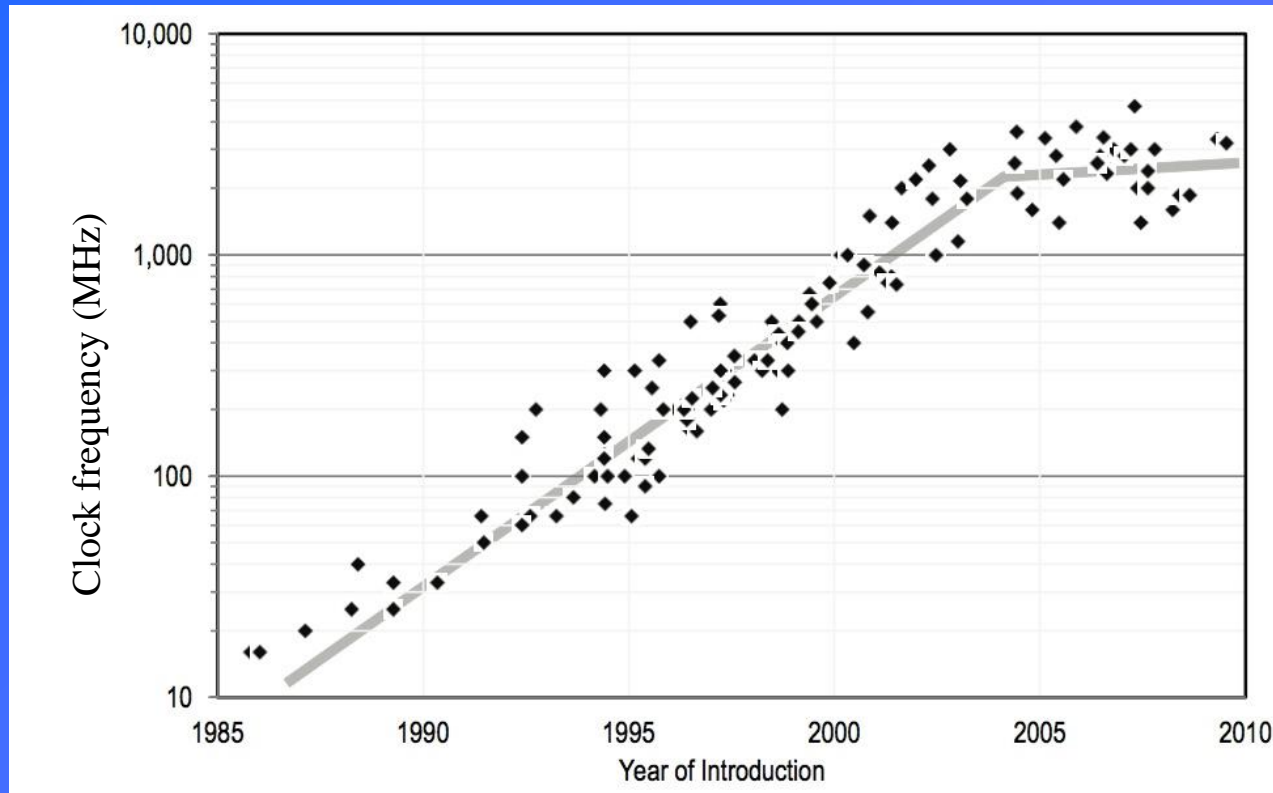
1985 : 1 watt/cm² 2016 : 145 watts/cm²



Source : THE FUTURE OF COMPUTING PERFORMANCE - Game Over or Next Level? Copyright 2011 by the National Academy of Sciences of the USA

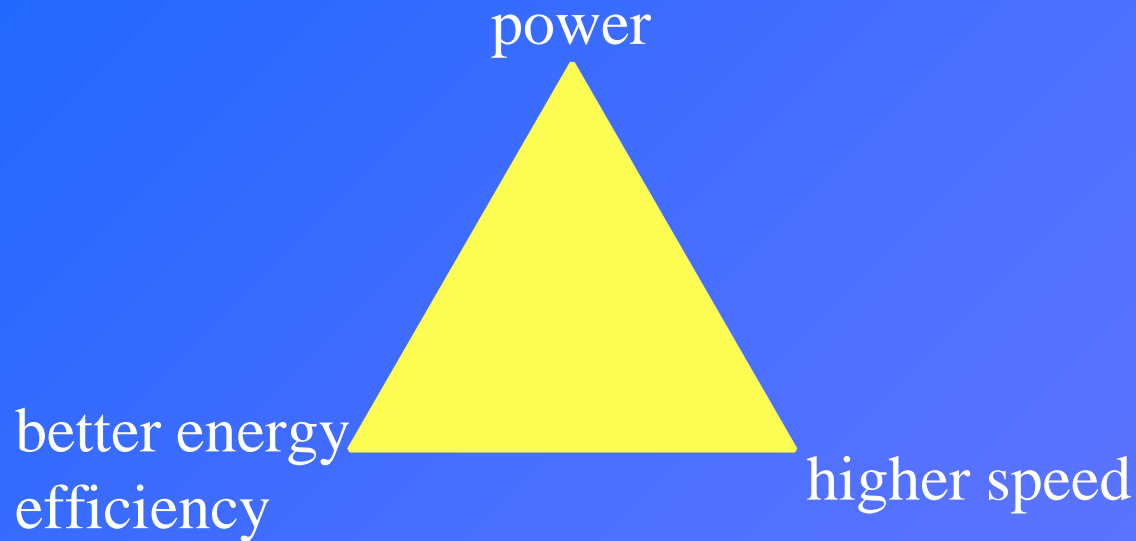
Clock frequencies of microprocessors

Clock frequencies of processors increased from about 10 MHz in 1985 to 3 GHz in 2005 : 40% increase of frequency each year for two decades.



Source : THE FUTURE OF COMPUTING PERFORMANCE - Game Over or Next Level? Copyright 2011 by the National Academy of Sciences of the USA

Metrics to compare technologies



Definitions

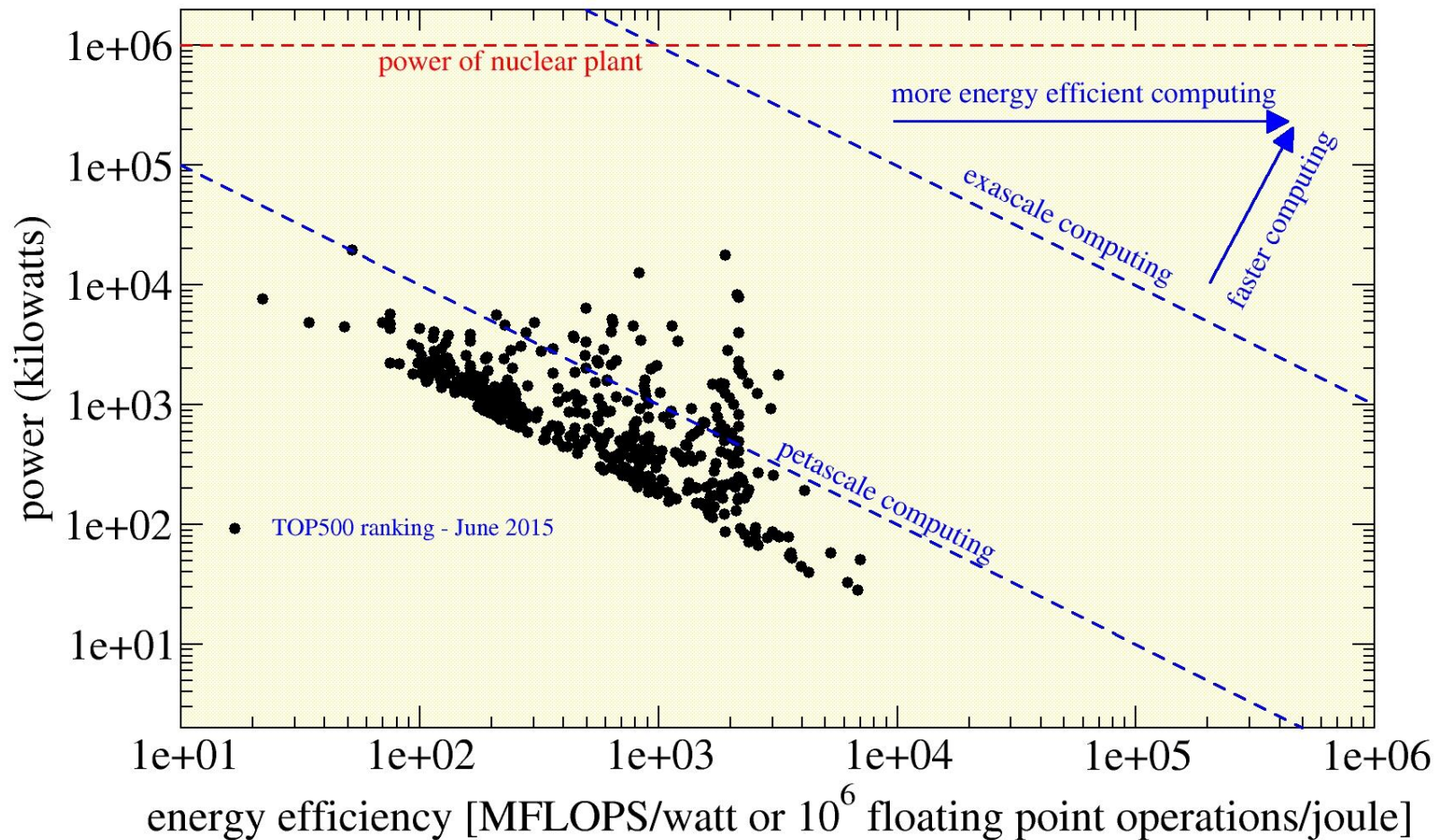
- FLOP : FLoating Point Operation
- energy efficiency = number of FLOPs per joule
- speed = number of FLOPs per second (speed means frequency of operation)

$$\text{speed} = \text{power} * \text{energy efficiency}$$

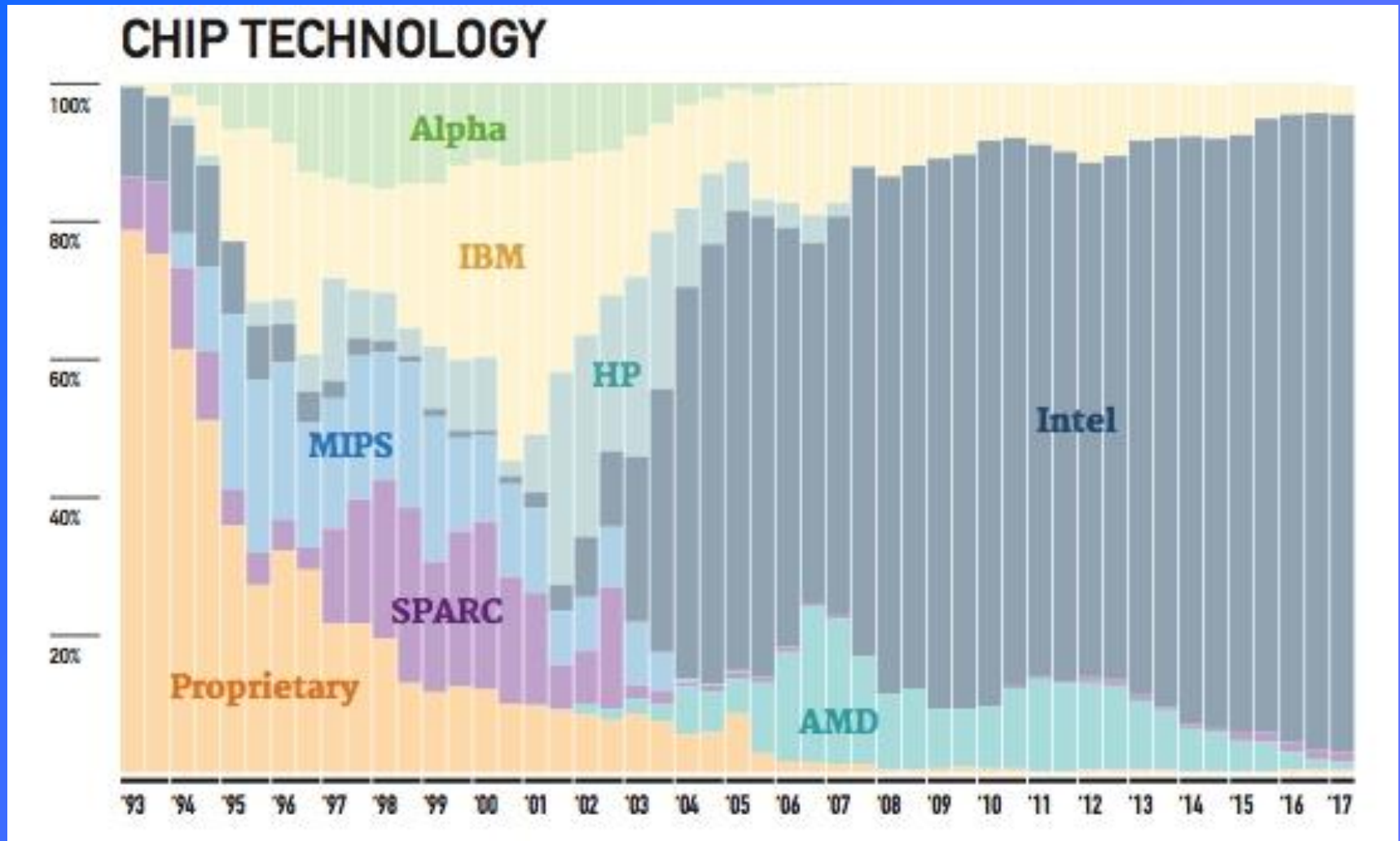
Current status of semiconductor high-end computing

Performance and power of high-end computing

September 2015 - total power consumption : 0.61 GW



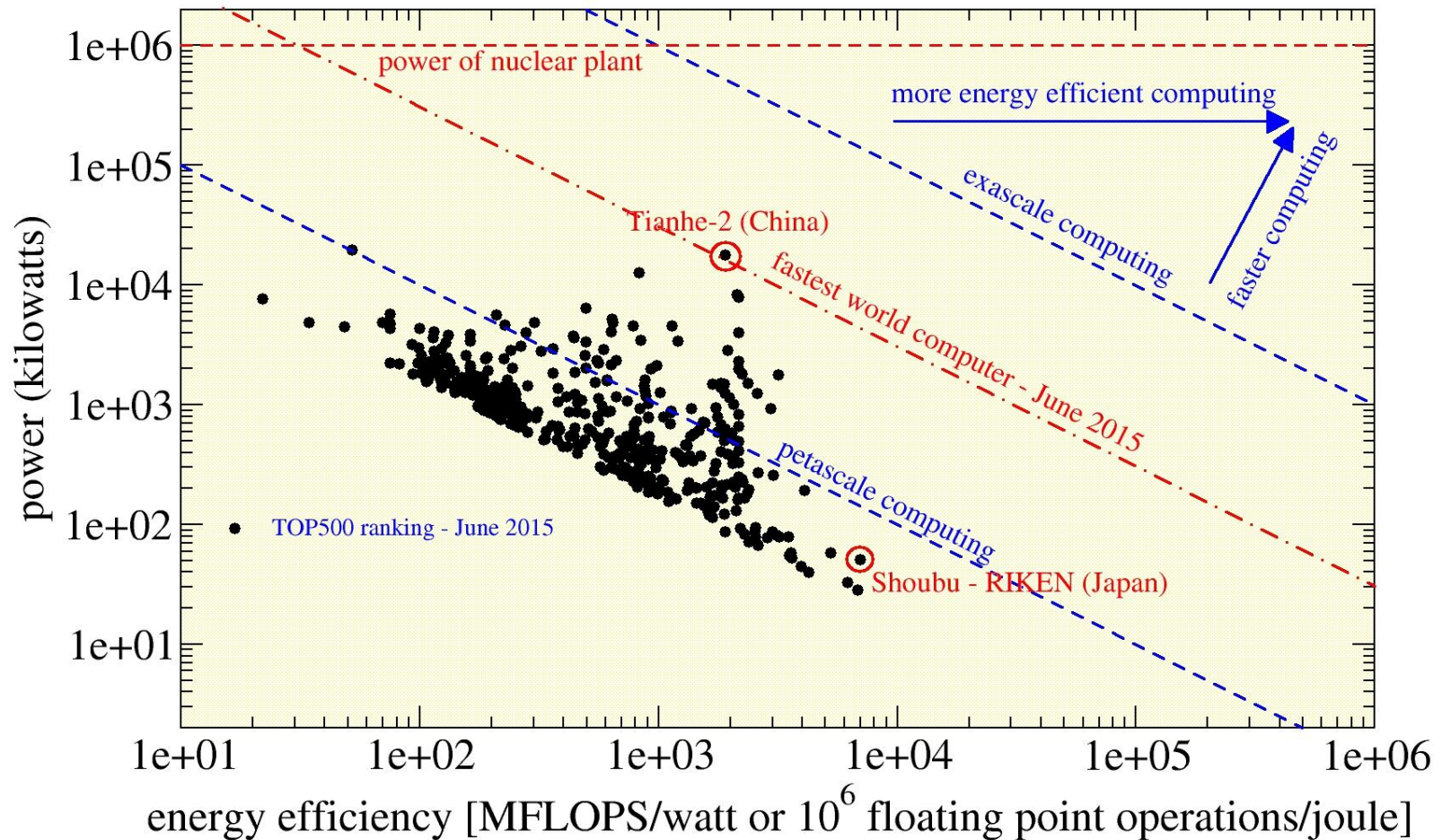
Semiconductor chips manufacturers

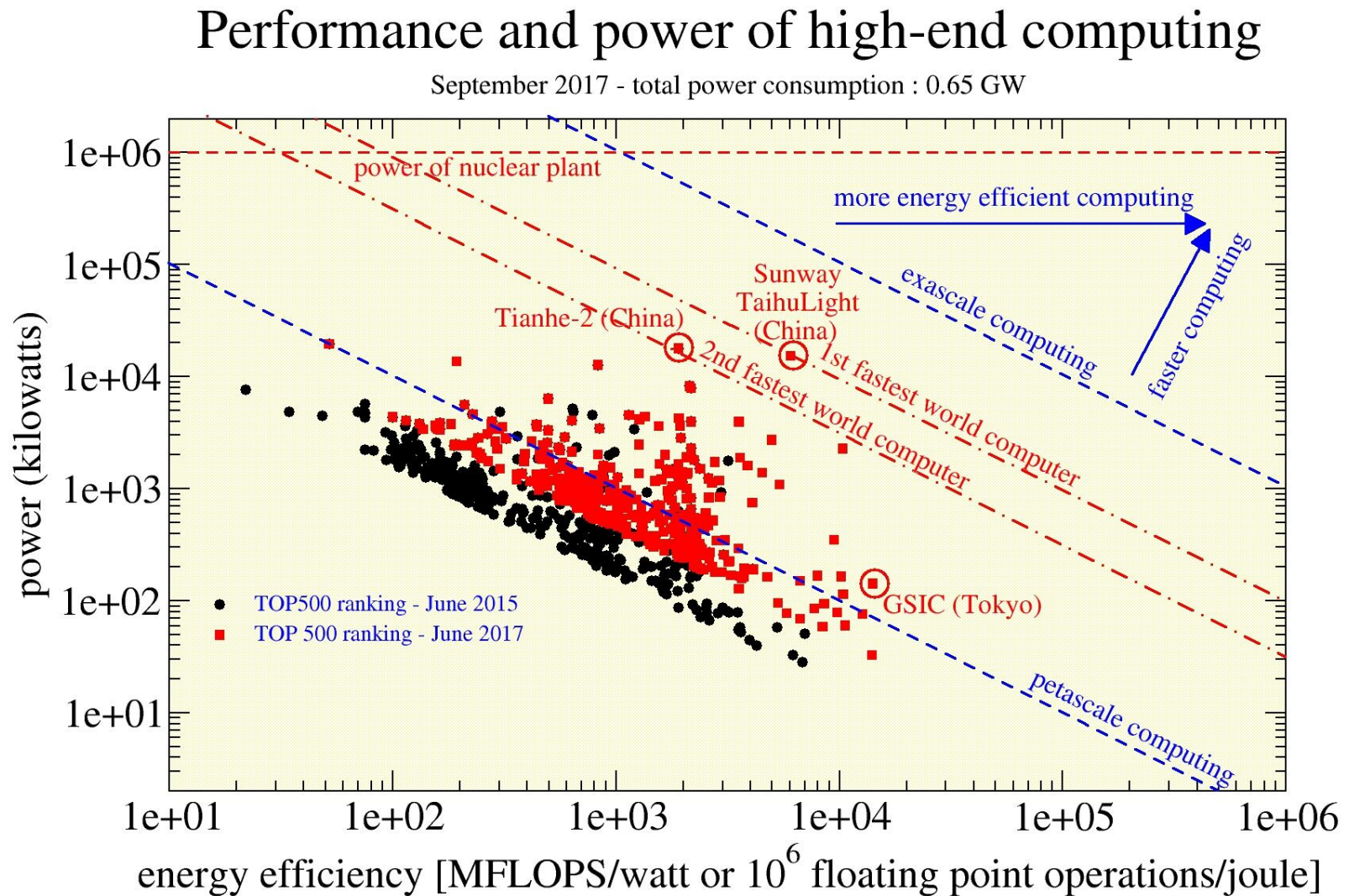


Current status of semiconductor high-end computing

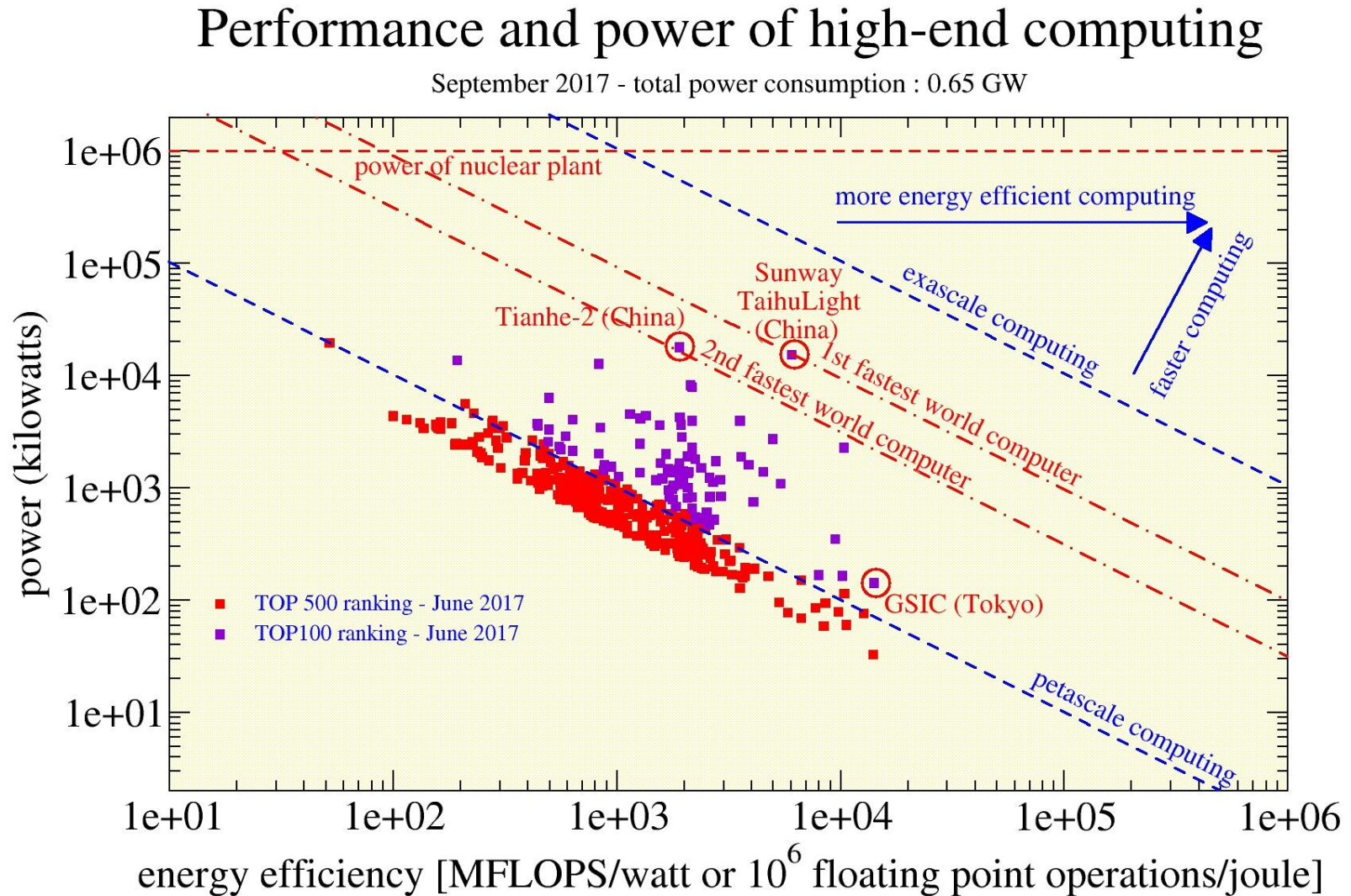
Performance and power of high-end computing

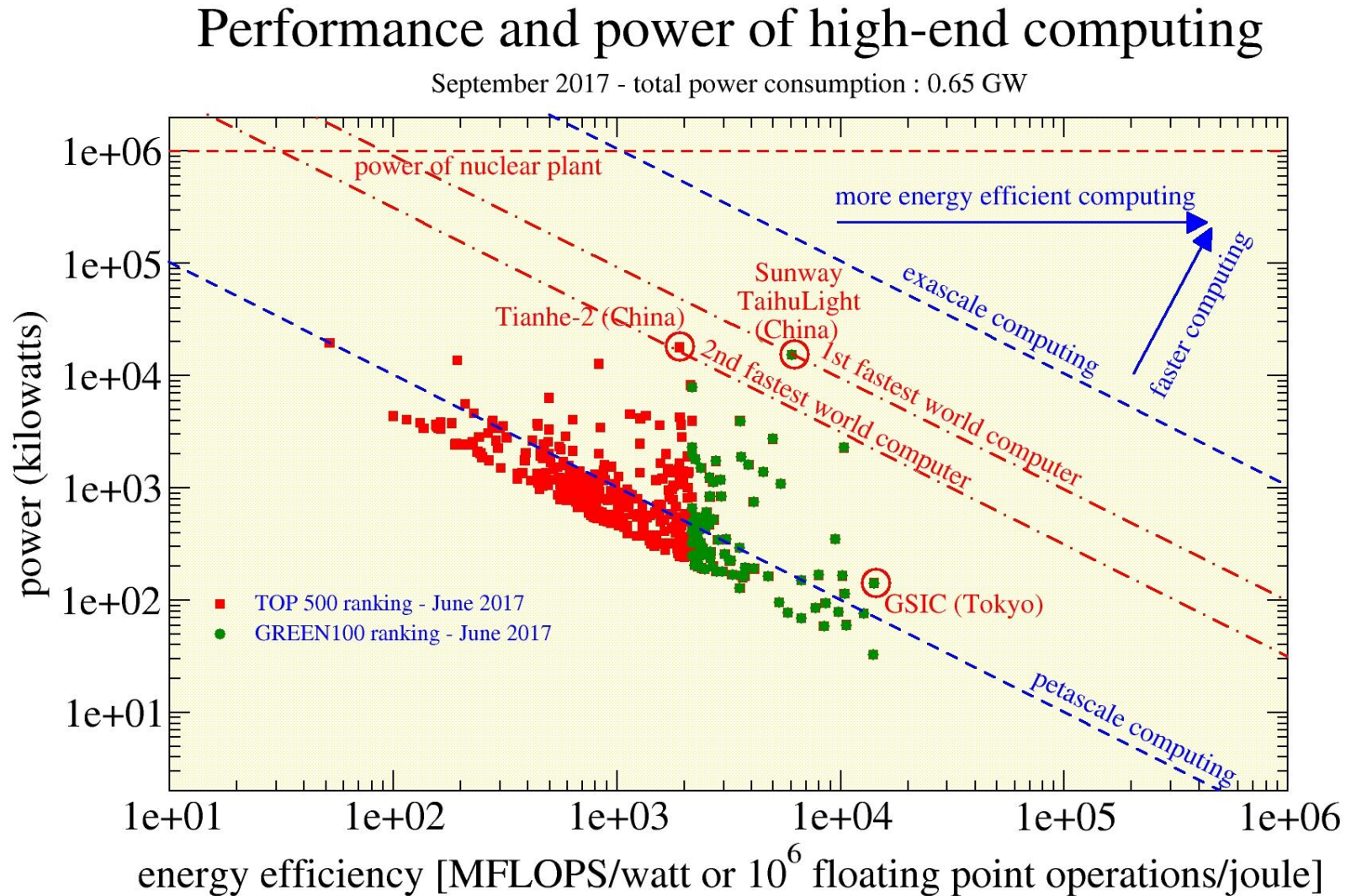
September 2015 - total power consumption : 0.61 GW





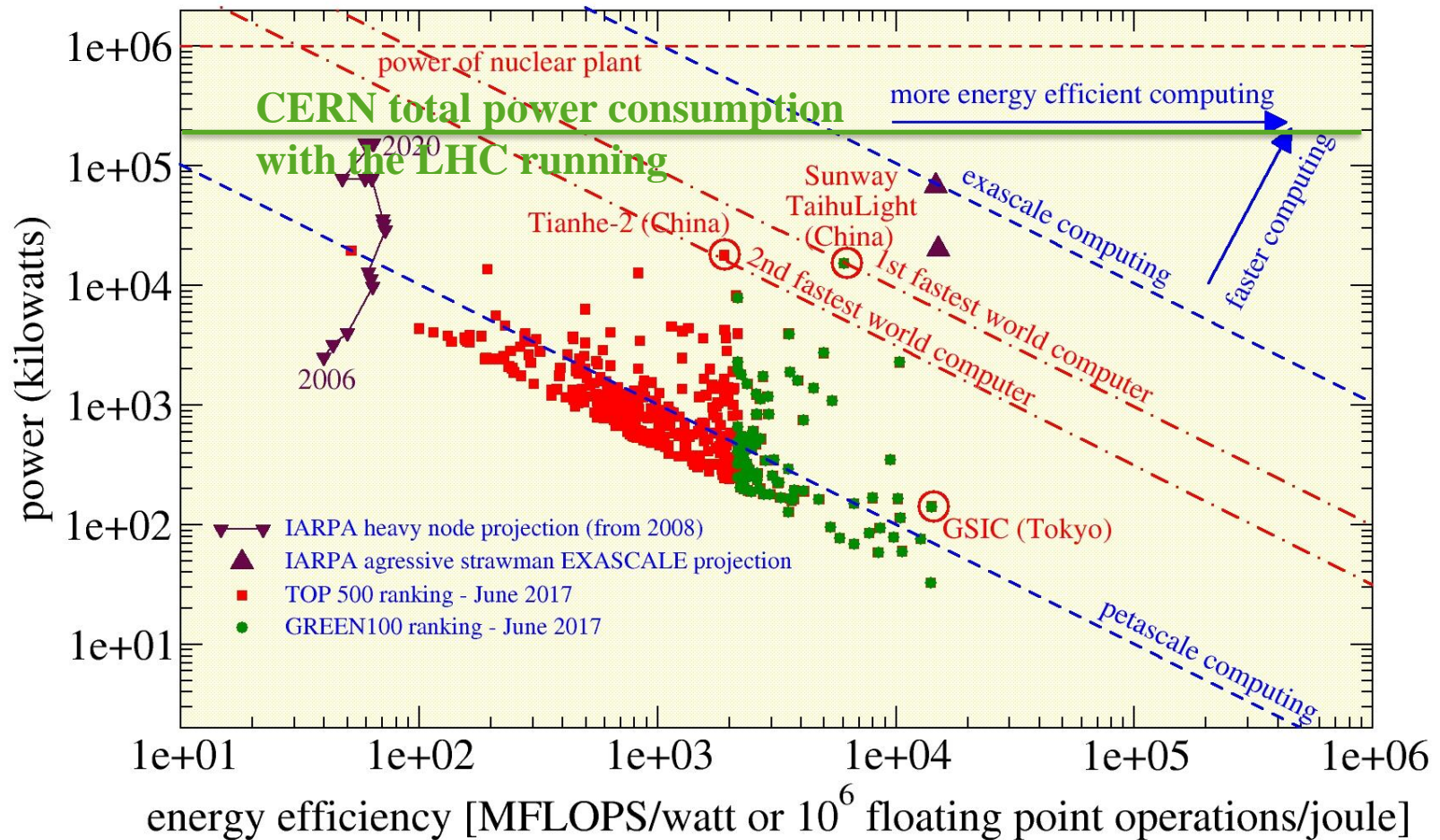
Current status of semiconductor high-end computing





Performance and power of high-end computing

September 2017 - total power consumption : 0.65 GW



Semiconductors : energy-delay product

Semiconductors : the dynamic power is the limiting quantity :

$$P_{dd} = C V_{dd}^2 f$$

- V_{dd} is the supply voltage
- C is the intrinsic gate capacitance

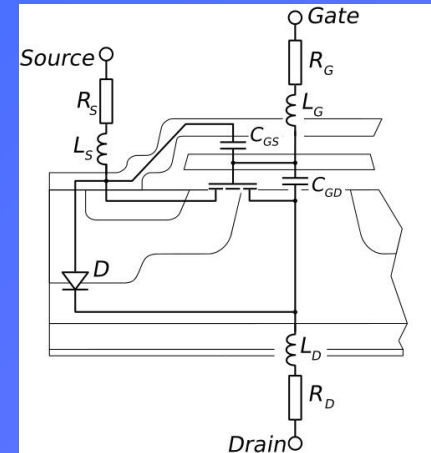
The intrinsic gate delay is : $\tau = \frac{C V_{dd}}{I_d}$

- I_d is the drain saturation current

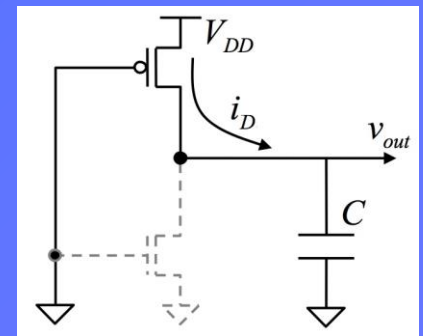
The energy-delay product (EDP) is :

$$EDP = \tau \frac{P_{dd}}{f} = \frac{C V_{dd}}{I_d} C V_{dd}^2 = \boxed{C^2} \boxed{\frac{V_{dd}^3}{I_d}}$$

energy-delay product (EDP) = 1/(energy efficiency² * power) = power/speed²



reduce transistor size
reduce supply voltage



Semiconductors : projection for the coming decade

Summary Table of ITRS Technology Trend Targets

Year of Production	2013	2015	2017	2019	2021	2023	2025	2028
Logic Industry "Node Name" Label	"16/14"	"10"	"7"	"5"	"3.5"	"2.5"	"1.8"	
Logic ½ Pitch (nm)	40	32	25	20	16	13	10	7
Flash ½ Pitch [2D] (nm)	18	15	13	11	9	8	8	8
DRAM ½ Pitch (nm)	28	24	20	17	14	12	10	7.7
FinFET Fin Half-pitch (new) (nm)	30	24	19	15	12	9.5	7.5	5.3
FinFET Fin Width (new) (nm)	7.6	7.2	6.8	6.4	6.1	5.7	5.4	5.0
6-t SRAM Cell Size(um ²) [@60f2]	0.096	0.061	0.038	0.024	0.015	0.010	0.0060	0.0030
MPU/ASIC HighPerf 4t NAND Gate Size(um ²)	0.248	0.157	0.099	0.062	0.039	0.025	0.018	0.009
4-input NAND Gate Density (Kgates/mm) [@155f2]	4.03E+03	6.37E+03	1.01E+04	1.61E+04	2.55E+04	4.05E+04	6.42E+04	1.28E+05
Flash Generations Label (bits per chip) (SLC/MLC)	64G /128G	128G /256G	256G / 512G	512G / 1T	512G / 1T	1T / 2T	2T / 4T	4T / 8T
Flash 3D Number of Layer targets (at relaxed Poly half pitch)	16-32	16-32	16-32	32-64	48-96	64-128	96-192	192-384
Flash 3D Layer half-pitch targets (nm)	64nm	54nm	45nm	30nm	28nm	27nm	25nm	22nm
DRAM Generations Label (bits per chip)	4G	8G	8G	16G	32G	32G	32G	32G
450mm Production High Volume Manufacturing Begins (100Kwspm)				2018				
Vdd (High Performance, high Vdd transistors)[**]	0.86	0.83	0.80	0.77	0.74	0.71	0.68	0.64
1/(CV/I) (1/psec) [**]	1.13	1.53	1.75	1.97	2.10	2.29	2.52	3.17
On-chip local clock MPU HP [at 4% CAGR]	5.50	5.95	6.44	6.96	7.53	8.14	8.8	9.9
Maximum number wiring levels [unchanged]	13	13	14	14	15	15	16	17
MPU High-Performance (HP) Printed Gate Length (GLpr) (nm) [**]	28	22	18	14	11	9	7	5
MPU High-Performance Physical Gate Length (GLph) (nm) [**]	20	17	14	12	10	8	7	5
ASIC/Low Standby Power (LP) Physical Gate Length (nm) (GLph)[**]	23	19	16	13	11	9	8	6

**** Note:** from the PIDS working group data; however, the calibration of Vdd, GLph, and I/CV is ongoing for improved targets in 2014 ITRS work

Source : International Technology Roadmap for Semiconductors – 2013 edition – Executive summary

Energy-delay product : projection

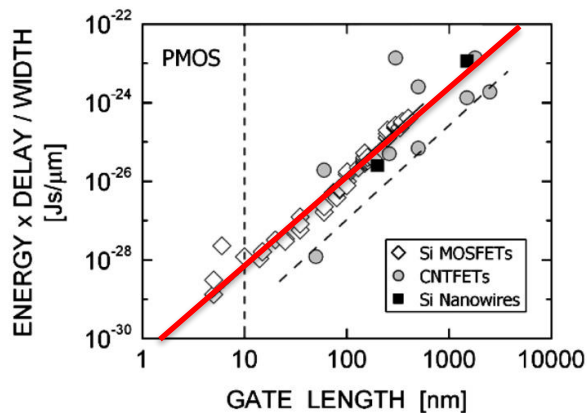


Fig. 6. Energy-delay product per device width versus transistor physical gate length of PMOS transistors.

Source : Robert Chau et al, IEEE Trans. Nanotechnology, Vol.. 4, No. 2, March 2005

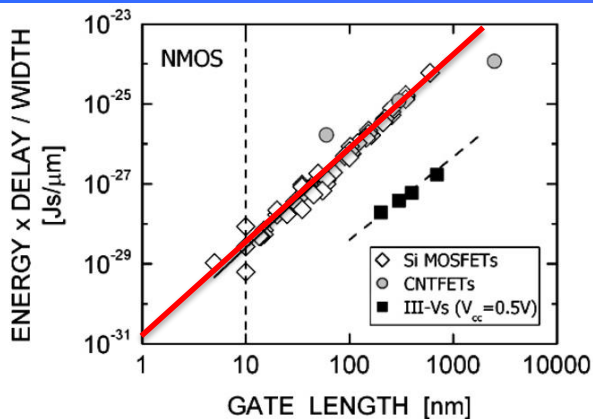


Fig. 7. Energy-delay product per device width versus transistor physical gate length of NMOS transistors.

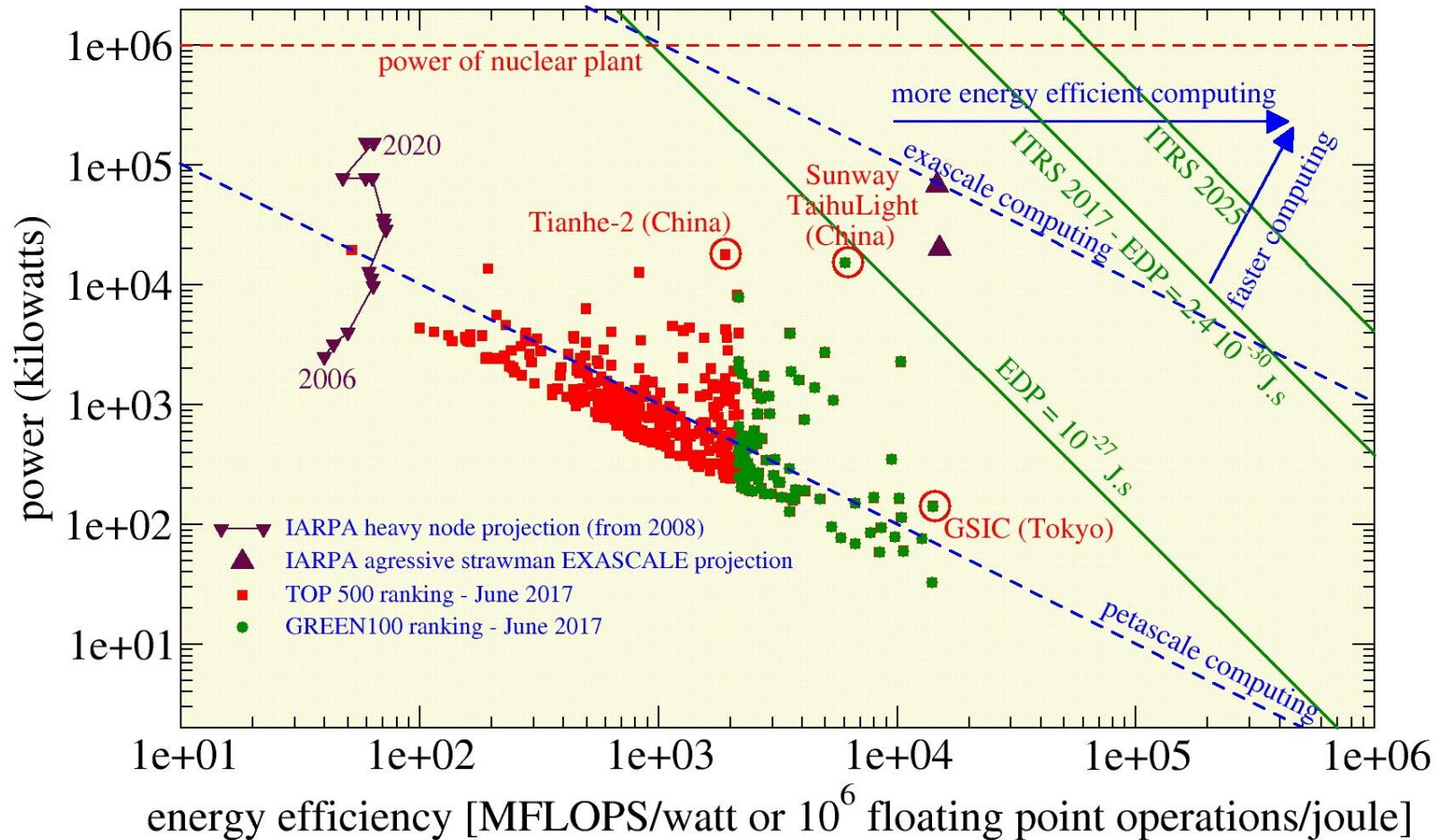
$$EDP[J \cdot s / \mu m] = 4 \cdot 10^{-31} \cdot gate\ length[nm]^{2.3}$$

$$EDP[J \cdot s] = 4 \cdot 10^{-34} \cdot gate\ length[nm]^{3.3}$$

Year	gate length	EDP (J.s/μm)	EDP (J.s)
2013	20	3,9E-28	7,9E-30
2015	17	2,7E-28	4,6E-30
2017	14	1,7E-28	2,4E-30
2019	12	1,2E-28	1,5E-30
2021	10	8,0E-29	8,0E-31
2023	8	4,8E-29	3,8E-31
2025	7	3,5E-29	2,5E-31
2028	5	1,6E-29	8,1E-32
2040	1	4,0E-31	4,0E-34

Performance and power of high-end computing

September 2017 - total power consumption : 0.65 GW

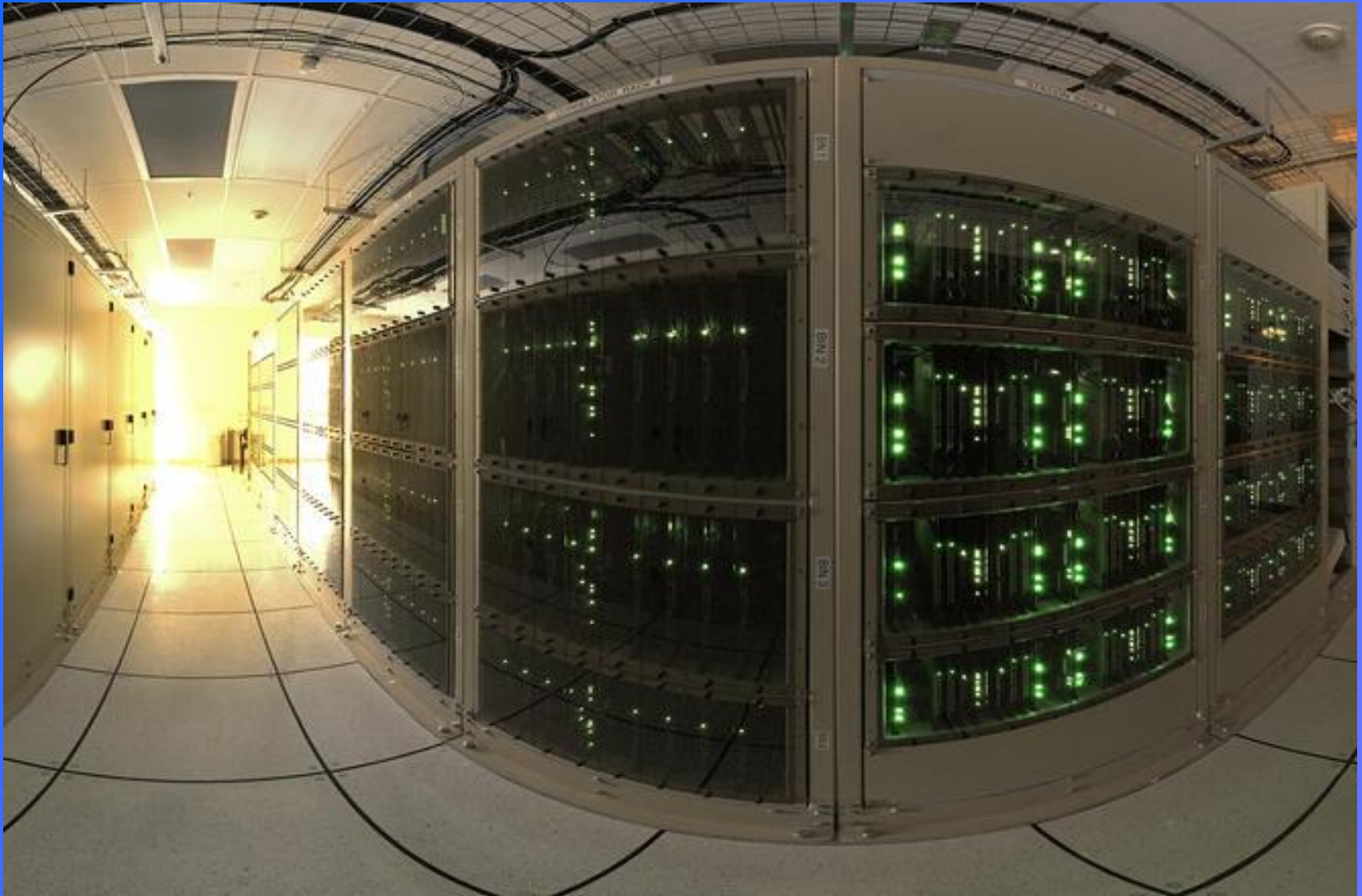


Supercomputers for astronomy



Atacama Large Millimeter Array (ALMA) - Source : ESO

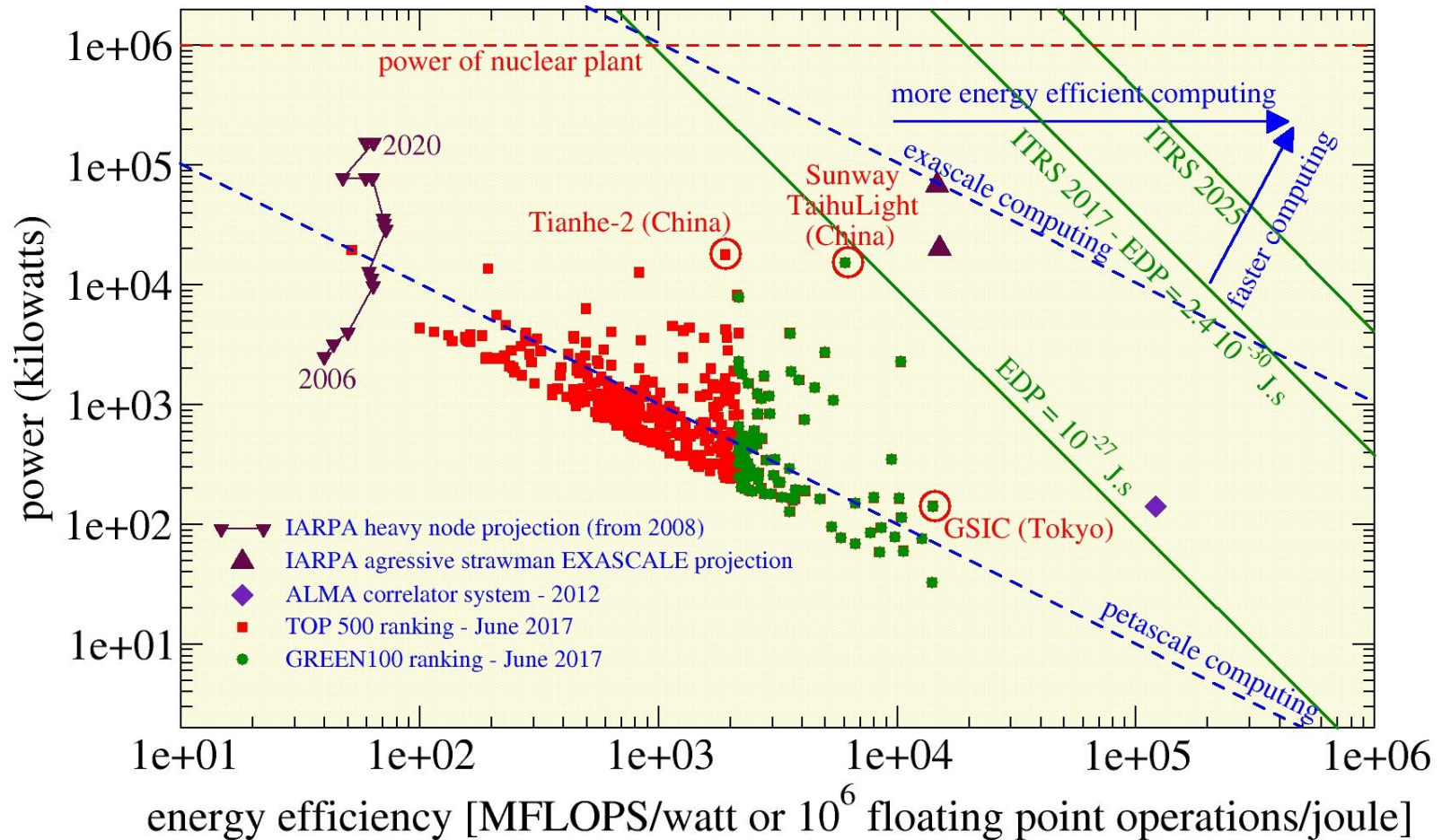
ALMA correlator



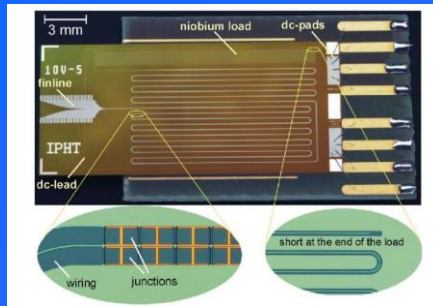
One of the four quadrants making up the ALMA correlator - Source : ESO

Performance and power of high-end computing

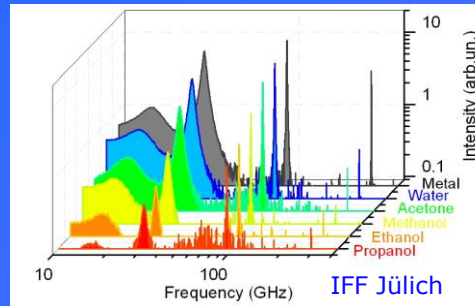
September 2017 - total power consumption : 0.65 GW



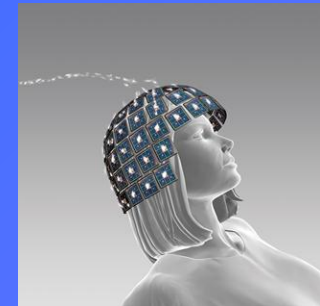
Superconducting computing with Josephson junctions



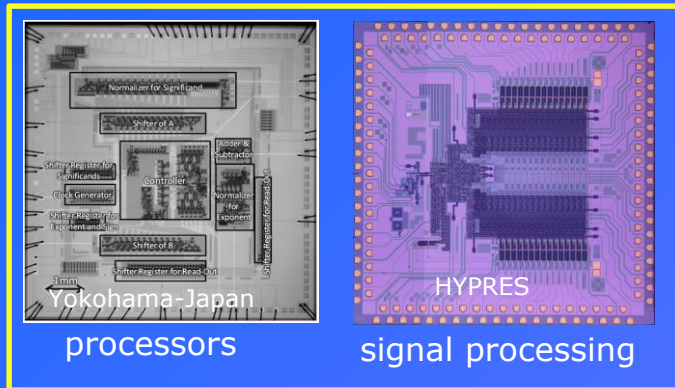
metrology



spectroscopy

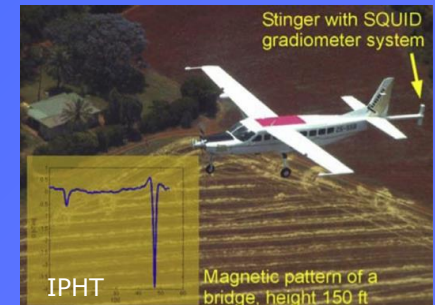


magneto-encephalography

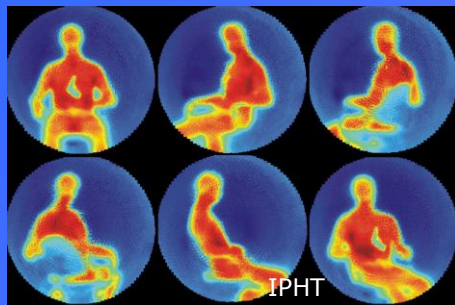


Josephson
junction

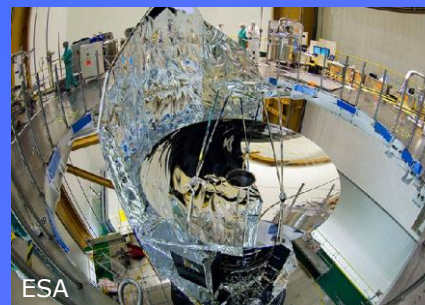
IPHT LEI 15.0kV X2,000 10µm WD 14.1mm



geophysics



security



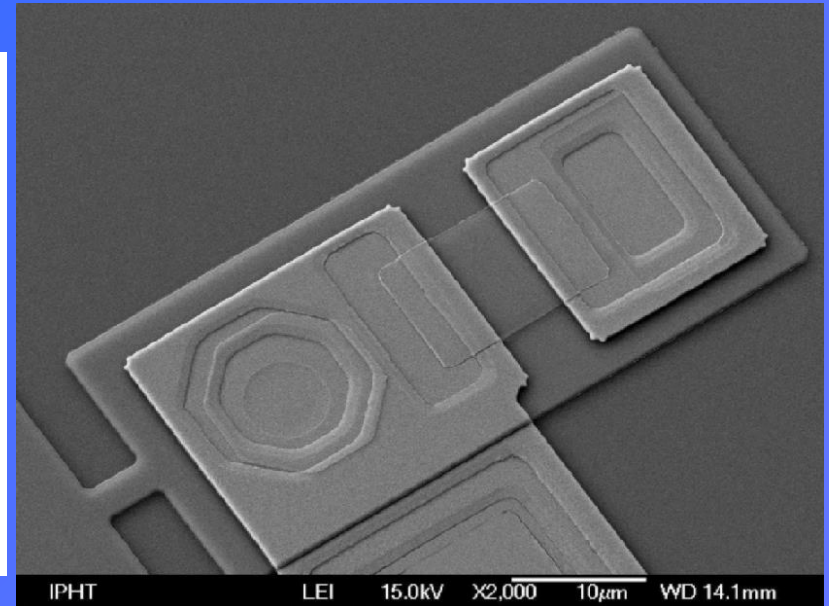
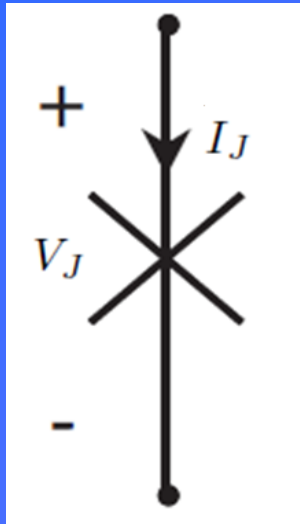
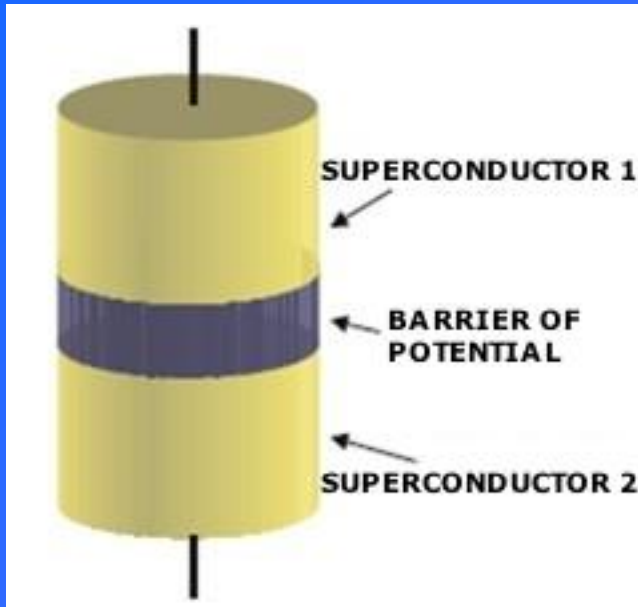
radio-astronomy



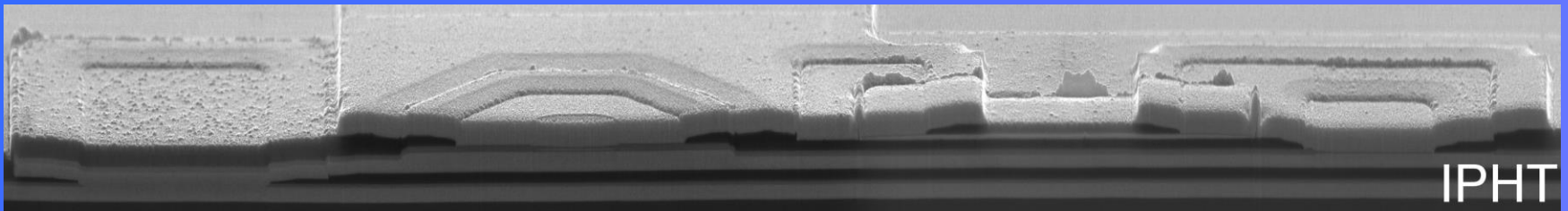
magnetic field imaging

The Josephson junction

The Josephson junction: **the** active element of superconductive electronics



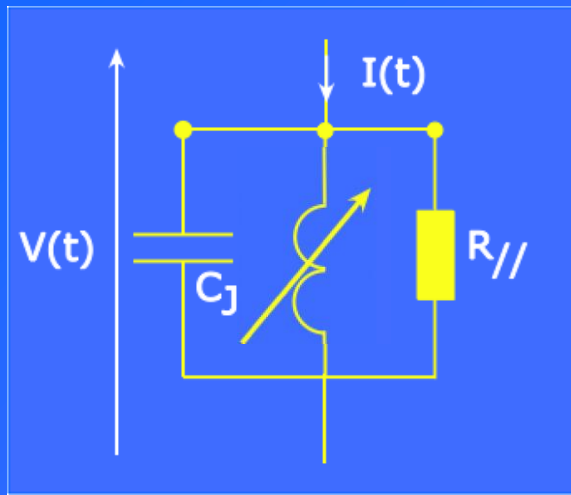
Most commonly used materials: Nb/Al-AlO_x/Nb @ 4.2 K



3 μm

Pictures from the FLUXONICS Foundry - IPHT Jena - Germany

Josephson junction electrodynamics

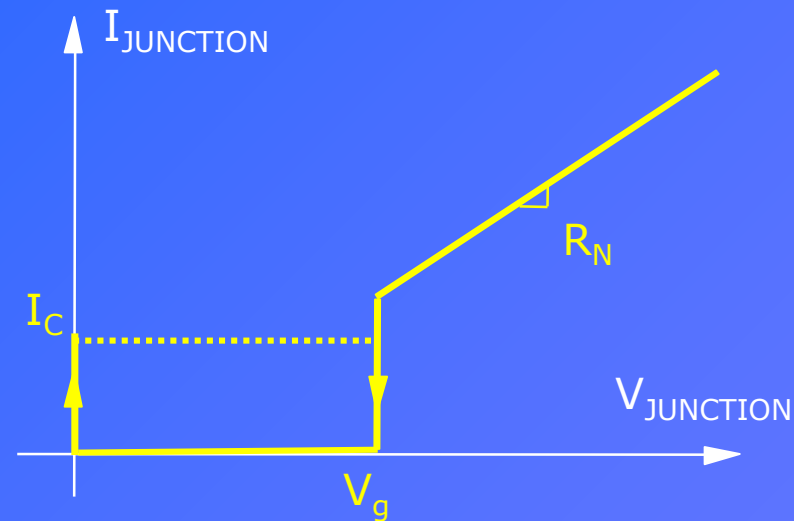


$$I(t) = I_C(t) + I_J(t) + I_R(t)$$

$$I_J(t) = I_c \sin \varphi(t)$$

$$I_C(t) = C_J \frac{\partial V(t)}{\partial t}$$

$$I_R(t) = \frac{V(t)}{R_J}$$



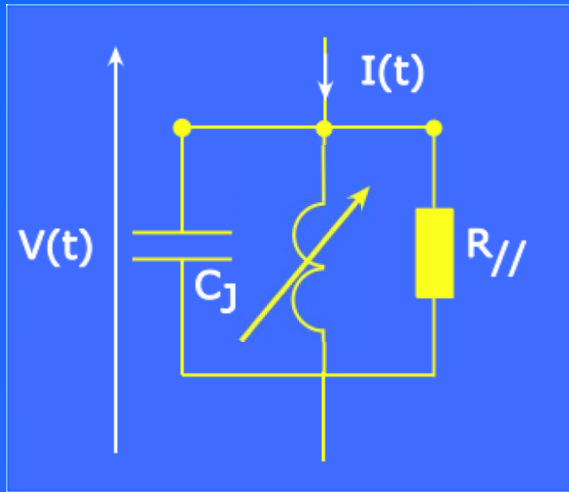
2nd Josephson equation (Faraday's law): $V(t) = \frac{\Phi_0}{2\pi} \frac{\partial \varphi(t)}{\partial t}$

$$\longrightarrow V(t) = \frac{\Phi_0}{2\pi} \left[\frac{1}{I_c \cos \varphi(t)} \frac{\partial I_J(t)}{\partial t} \right] = L_J \frac{\partial I_J(t)}{\partial t}$$

$$L_J = \frac{L_{J_0}}{\cos \varphi(t)} \text{ with } L_{J_0} = \frac{\Phi_0}{2\pi I_c} \quad \text{Josephson inductance}$$

\longrightarrow JJ = non-linear parallel RLC circuit

Josephson junction time constants



$$\frac{I(t)}{I_c} = L_{J_0} C_J \frac{\partial^2 \varphi(t)}{\partial t^2} + \frac{L_{J_0}}{R_{||}} \frac{\partial \varphi(t)}{\partial t} + \sin \varphi(t)$$

Anharmonic oscillator

Electrical approach

Physical approach
(BCS theory)

Plasma period
of the L-C circuit :

$$\tau_p = 2\pi \sqrt{L_{J0} C_J}$$

$$\tau_p = \sqrt{\frac{2\pi \Phi_0 C_S}{j_c}}$$

L-R circuit time constant :

$$\tau_c = \frac{L_{J0}}{R_{||}}$$

$$\tau_c = \frac{2\Phi_0}{\pi^2 V_g}$$

Nb: 0.15 ps
NbN: 0.07 ps

R-C circuit time constant :

$$\tau_{RC} = R_{||} C_J$$

$$\tau_{RC} = \frac{\pi V_g C_S}{4 j_c}$$

Minimizing the switching time

$$\tau_p = 2\pi \sqrt{L_{J0} C_J}$$

$$\tau_{LR} = \frac{L_{J0}}{R_{//}}$$

$$\tau_{RC} = R_{//} C_J$$

L-C circuit plasma period

L-R circuit time constant

R-C circuit time constant

McCumber parameter defined by:

$$\beta_c = \frac{\tau_{RC}}{\tau_{LR}} = \frac{R_{//}^2 C_J}{L_{J0}} = \frac{2\pi R_{//}^2 C_J I_c}{\Phi_0}$$

Minimum switching time obtained for :

$$\tau_{RC} = \tau_{LR} \left(= \frac{\tau_p}{2\pi} \right) : \beta_c = 1$$

New time constant

$$\tau_0 = \sqrt{\frac{\Phi_0 C_S}{2\pi j_c}} = \frac{\Phi_0}{2\pi R_{shunt} I_c}$$

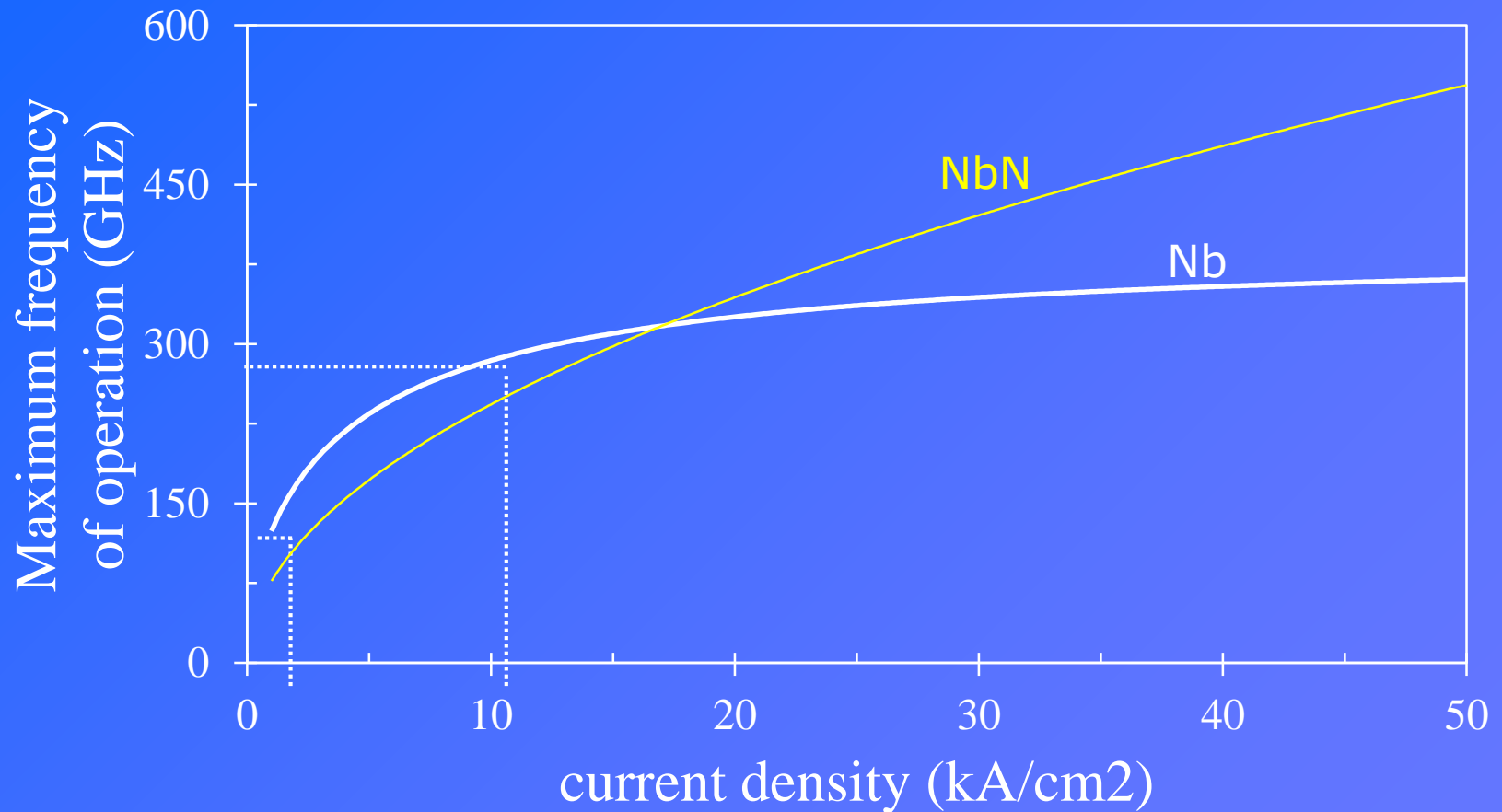
$$\tau_0 (ps) \approx \frac{1}{\pi V_c (mV)} \text{ with } V_c = R_{shunt} I_c$$

$$R_{//} \gg R_{shunt}$$

Criteria: $f_{\max} = 1/(2\pi\tau_0)$

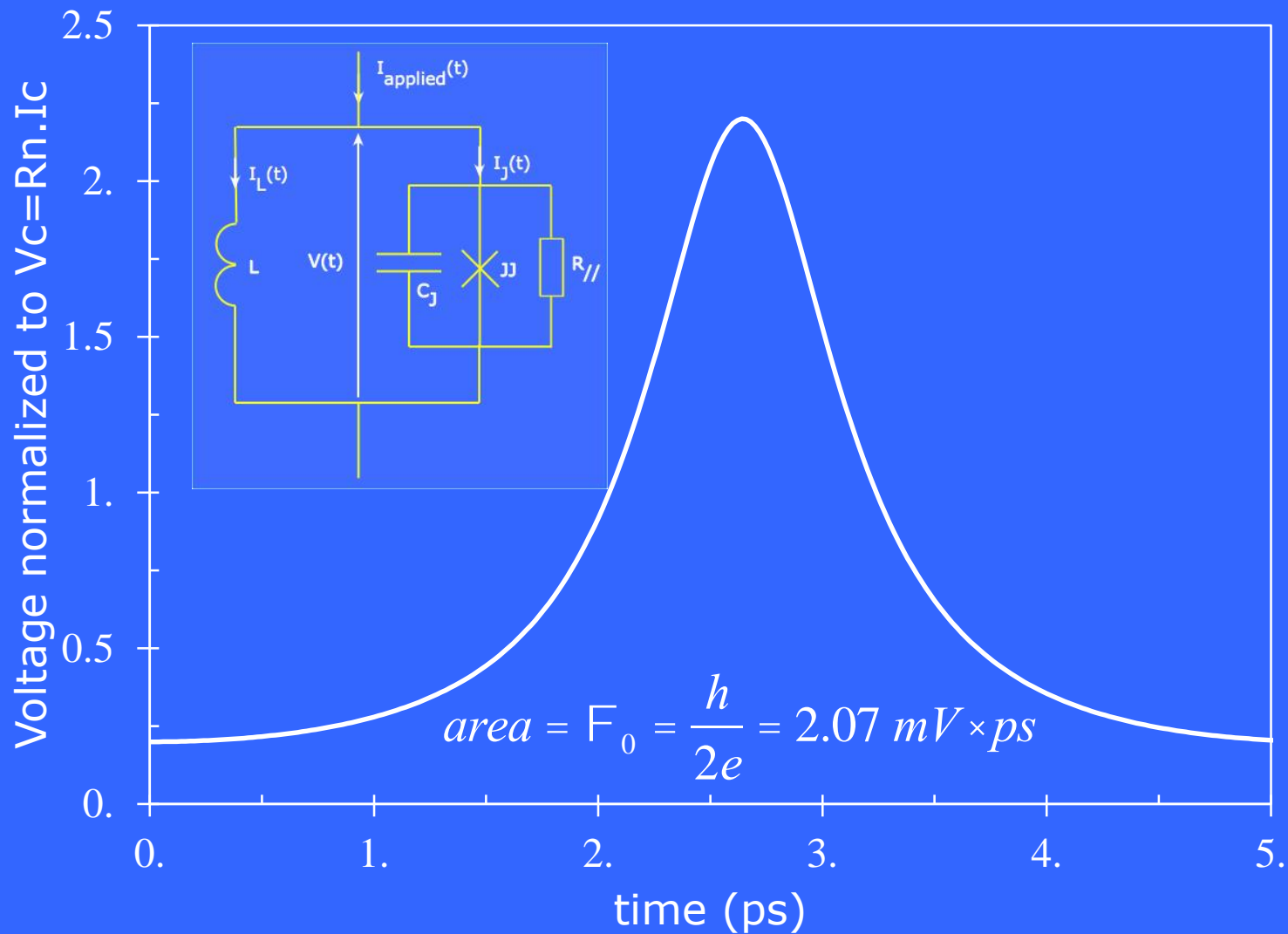
$$f_{\max} (\text{GHz}) = 500 \times V_c (\text{mV})$$

Maximum frequency of operation



Valid for externally-shunted SIS junctions

Rapid Single Flux Quantum (RSFQ) Logic



Superconducting electronics energy-delay product

$$EDP = \tau_0 I_c \Phi_0 = \frac{\Phi_0^2}{2\pi R_{shunt}} = \frac{6 \cdot 10^{-31}}{R_{shunt} [\Omega]} J \cdot s$$

The EDP does not depend on the size of devices for externally-shunted junctions.

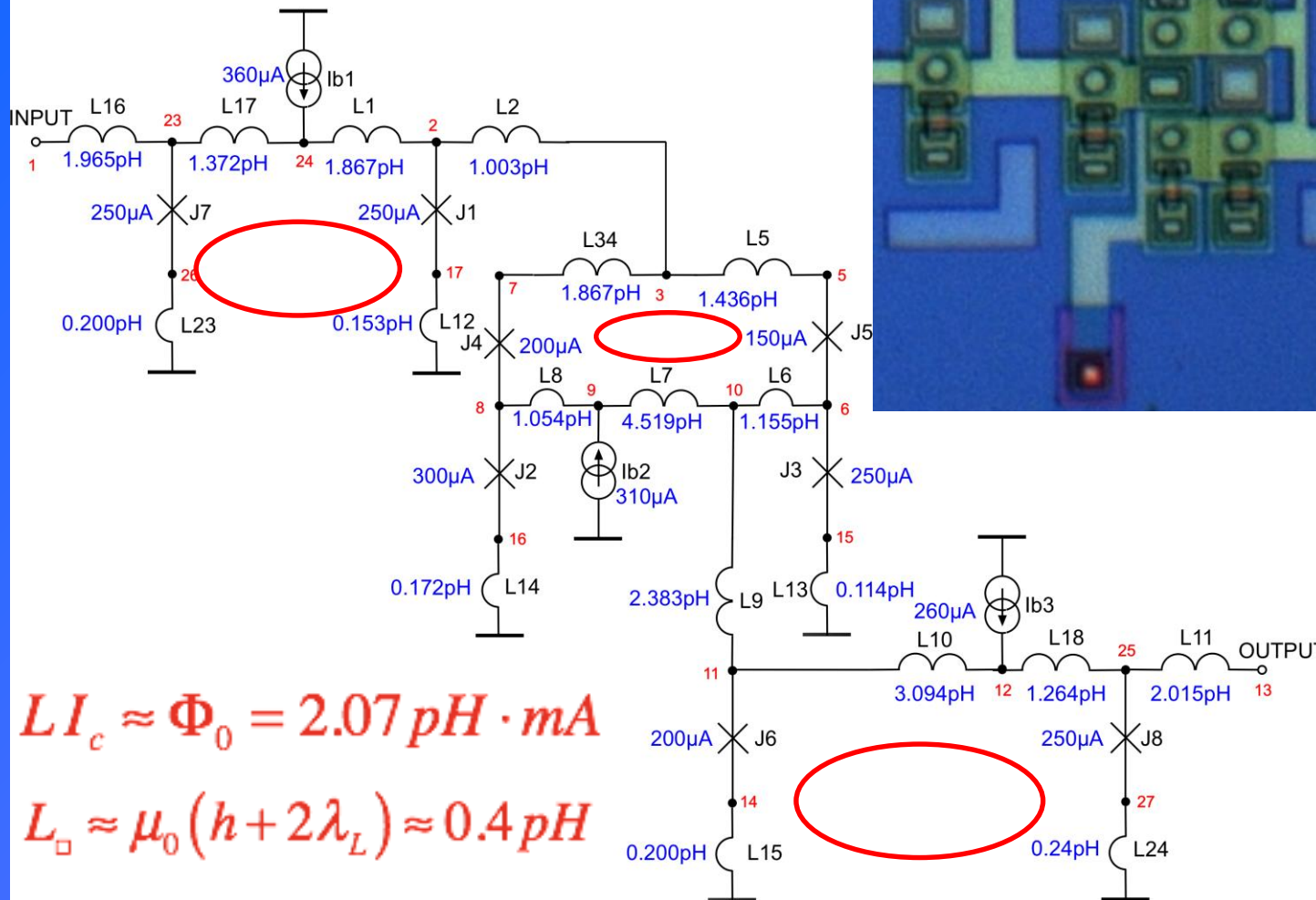
The EDP depends on the junction area for self-shunted junctions :

$$EDP \propto \frac{\Phi_0^2 A_{JJ}}{2\pi} \approx 10^{-30} \cdot A_{JJ} (\mu m^2) J \cdot s$$

$$EDP(semiconductors) = \frac{C^2 V_{dd}^3}{I_d}$$

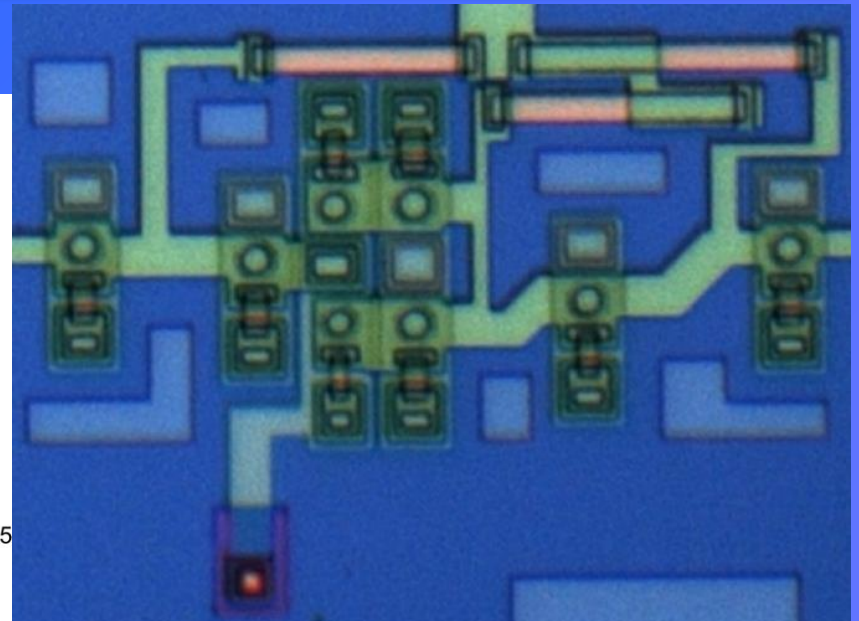
SFQ cells

FLUXONICS FOUNDRY - 2012 - TFF with parasitics



$$LI_c \approx \Phi_0 = 2.07 \text{ pH} \cdot \text{mA}$$

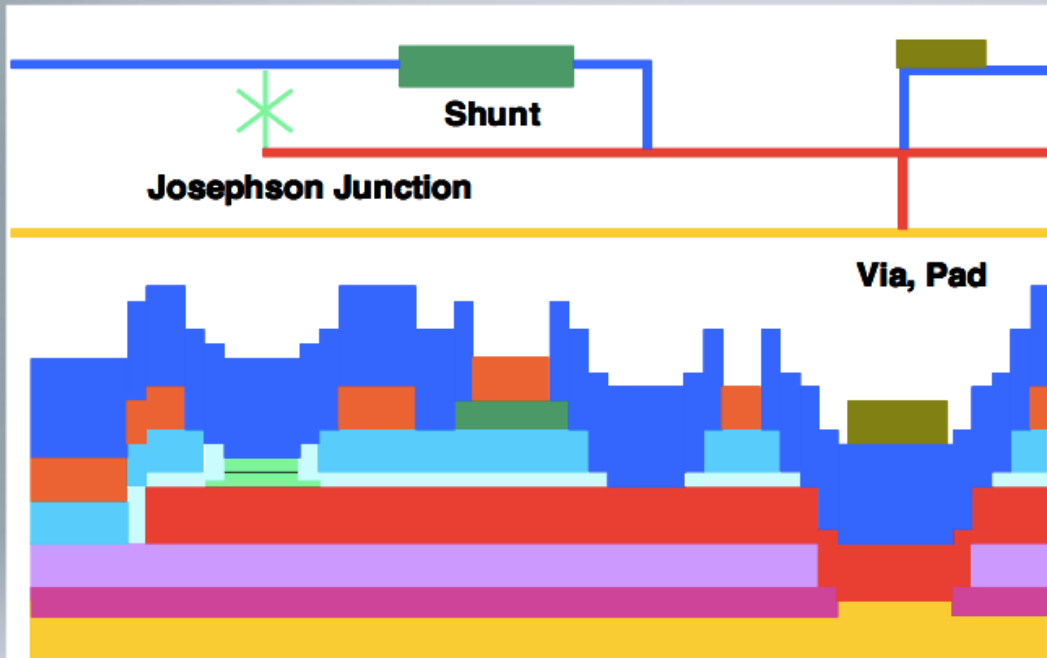
$$L_{\square} \approx \mu_0 (h + 2\lambda_L) \approx 0.4 \text{ pH}$$



www.FLUXONICS.eu

RSFQ process – cross section

iphtjena



Layer	Thickness	Material
R2	50 nm	Au
M2	350 nm	Nb
I2	150 nm	SiO
R1	80 nm	Mo
I1B	150 nm	SiO
I1A	70 nm	Nb ₂ O ₅
T1	60 nm 12 nm 30 nm	Nb Al ₂ O ₃ Nb
M1	250 nm	Nb
I0B	200 nm	SiO
I0A	50 nm	Nb ₂ O ₅
M0	200 nm	Nb



Superconducting digital electronics foundry process

PROCESS	Current density [kA/cm ²]	minimum area [μm ²]	Maximum integration	Maximum frequency [GHz]
Hypres #03-10-45	0.03 1.0 4.5	~ 3.14	15,000	80 GHz RnIc=1.3mV @ 4.5 kA/cm ²
Hypres #S45/100/200	0.1 1 4.5 10 20 30	~ 0.4	10,000	200 GHz @ 30 kA/cm ²
MIT Lincoln Lab SFQx	10 20 50	~ 0.06	~ 800,000	240 GHz RnIc=2.17 mV @50 kA/cm ²
ADP2	10	1.0	1100 JJ/mm ²	80 GHz
STP2	2.5 - 20	0.25 - 4.0	100 JJ/mm ² - > 2,000 JJ/mm ²	30 GHz - 150 GHz
HSTP	10	1.0	70,000	80 GHz
Fluxonics standard	1	12.5	100 JJ/mm ²	40 GHz RnIc =0.256 mV
INRiM SNIS	up to 100	25	1,000 JJ/mm ²	300 GHz RnIc =0.1mV - 0.7mV
NIST Nb/Nbx Si1-x/Nb	up to 110	?	70,000	300 GHz
INRiM SNIS 3D FIB	up to 100	0.1	10,000 JJ/mm ²	300 GHz RnIc=0.1mV - 0.7mV

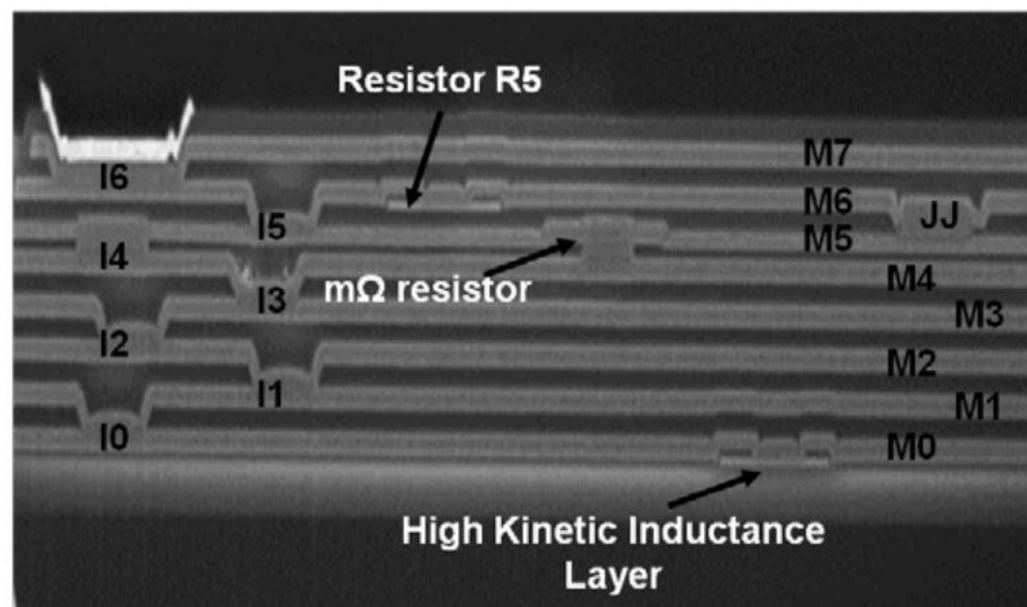
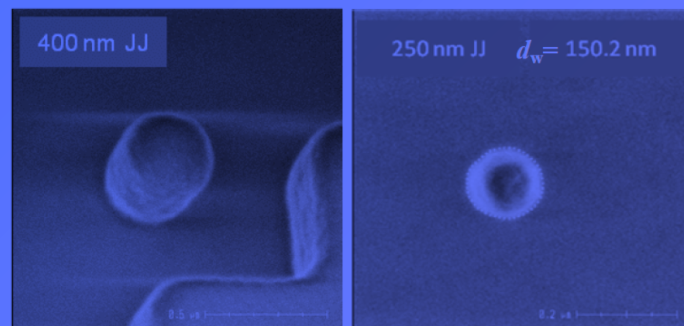
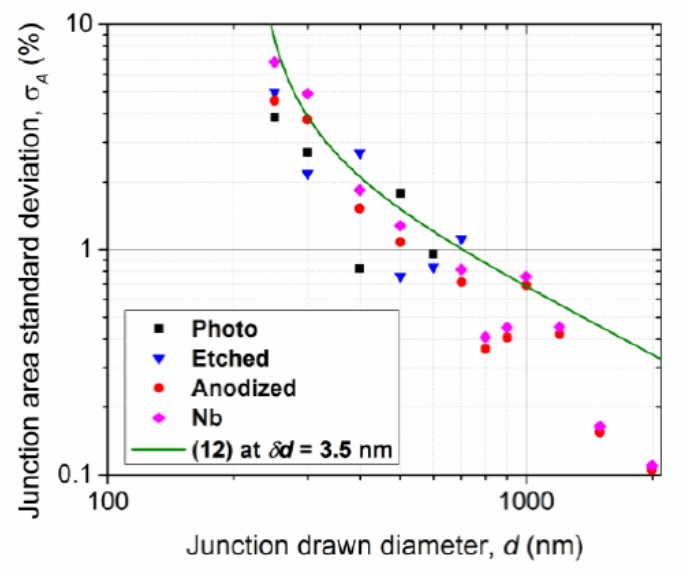


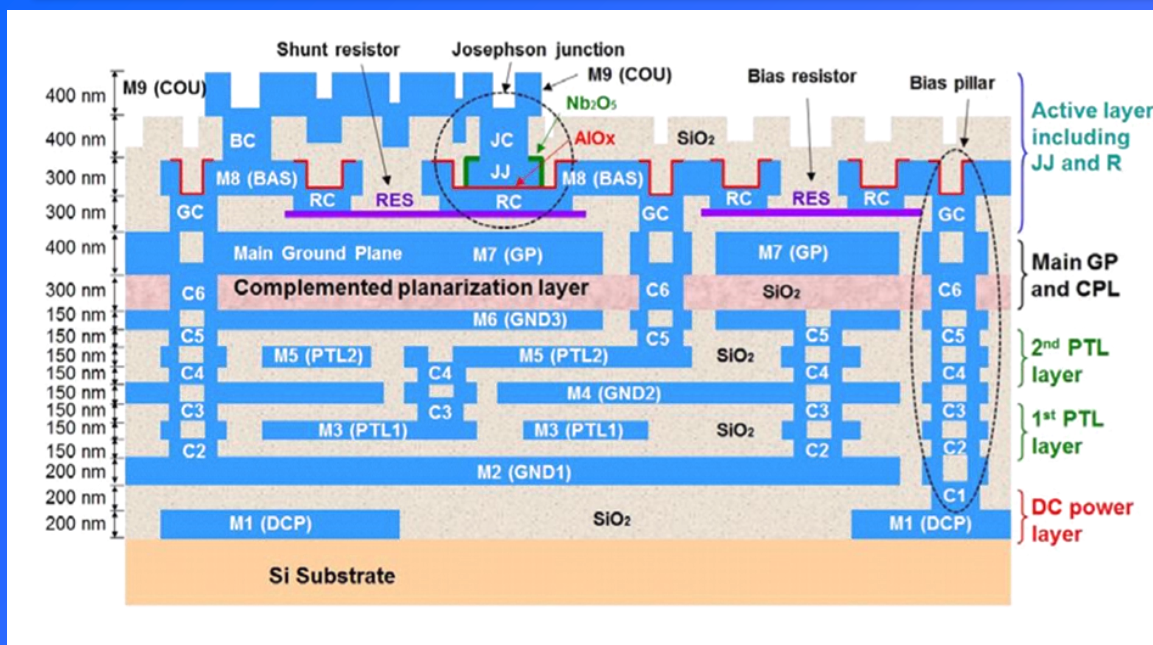
Fig. 10. Cross section of a wafer fabricated by the SFQ5ee process. The labels of metal layers and vias are the same as in Table I. New features of the SFQ5ee are shown: a high kinetic inductance layer under M0 and a layer of mΩ-range resistors between M4 and M5 layers.



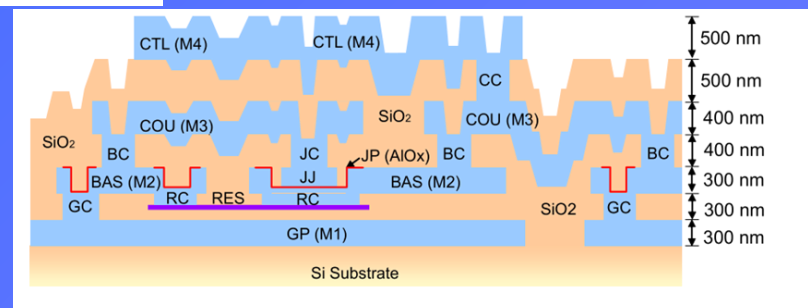
JJ count \approx 800,000 JJ/chip

S. K. Tolpygo et al, "Fabrication Process and Properties of Fully- Planarized Deep-Submicron Nb/Al-AIOx/Nb Josephson Junctions for VLSI Circuits," IEEE TAS 2015

S.K. Tolpygo et al., "Developments towards a 250-nm, fully planarized fabrication process with ten superconducting layers and self-shunted Josephson junctions," arxiv_1704.07683 (20017); *IEEE Trans. Appl. Supercond.* to be published.



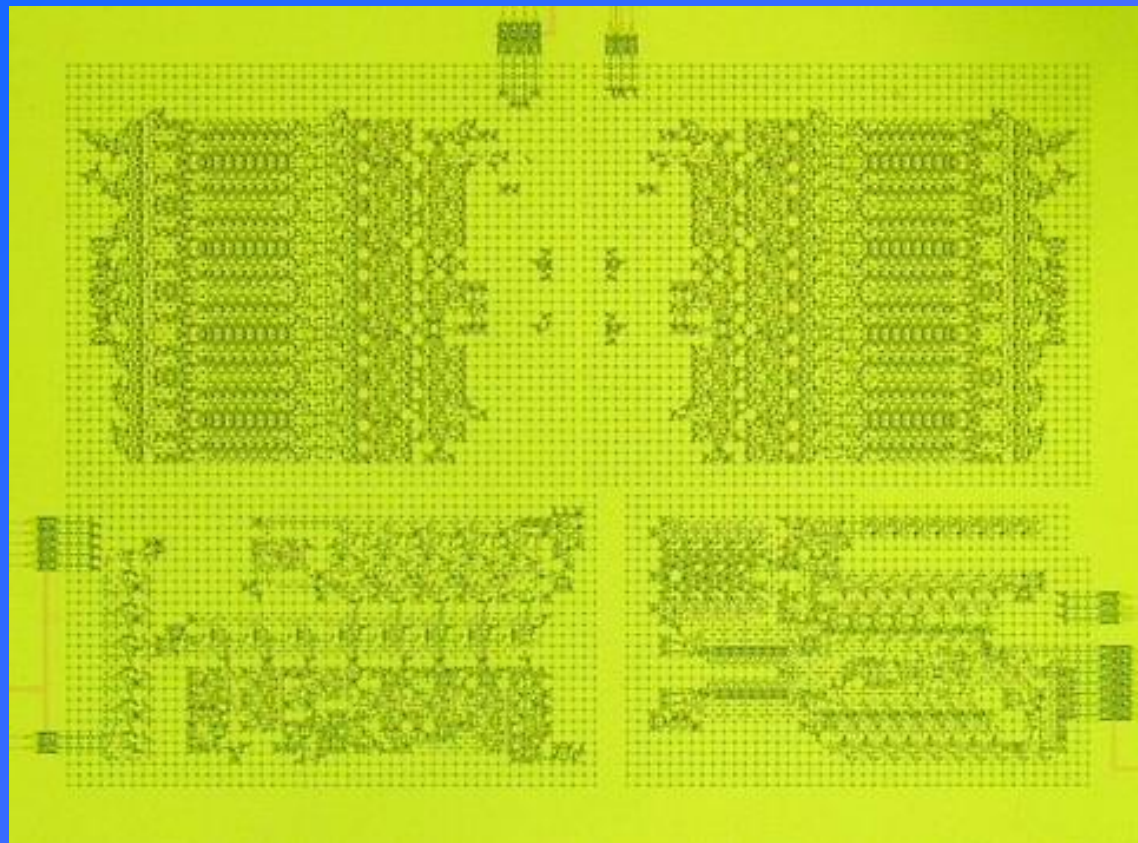
ADP 2 process (9 metal levels - 10kA/cm²)



STP2 process (4 metal levels - 2.5 or 20kA/cm²)

S. Nagasawa et al., "Nb 9-Layer Fabrication Process for Superconducting Large-Scale SFQ Circuits and Its Process Evaluation," IEICE 2014

S. Nagasawa, T. Satoh, and M. Hidaka, "Uniformity and Reproducibility of Submicron 20kA/cm² Nb/AIOx/Nb Josephson Junction Process," ISEC 2015



COREe2 (2017)

10655 JJs

500 MIPS

2.4 mW

210 GIPS/W

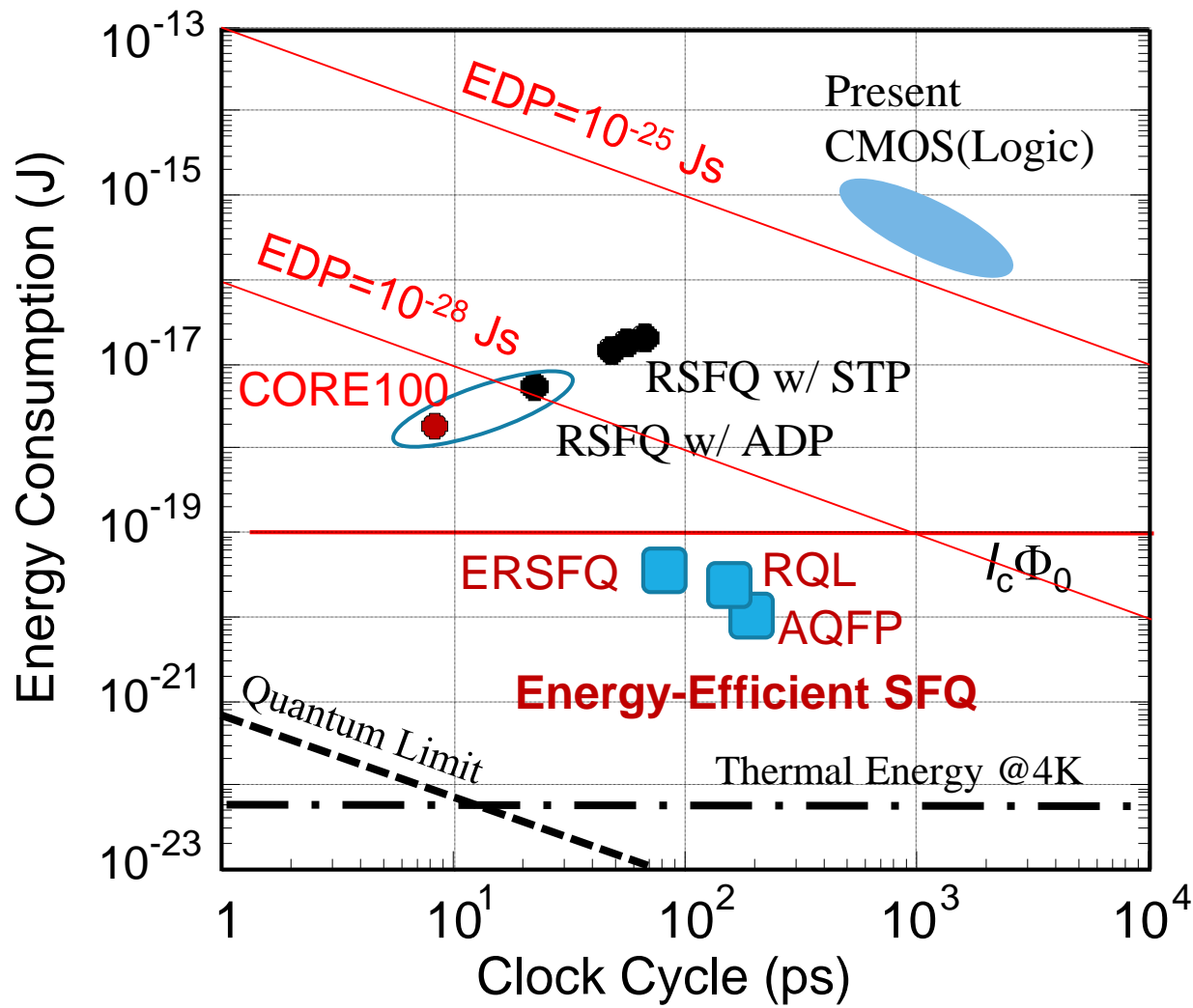
Programs Executed

50 GHz

Memory Embedded

Courtesy : Prof. Akira FUJIMAKI - Nagoya University

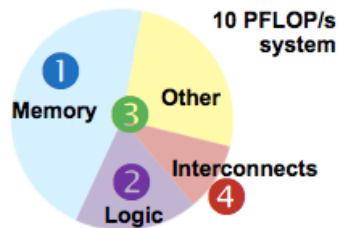
Energetic considerations



Courtesy : Prof. Akira FUJIMAKI - Nagoya University

Cryogenic Computer Complexity (C3) project - IARPA

Performance (PFLOP/s):	1	10	100	1,000
Power budget (@ 4 K)	1.5 W	10 W	100 W	1,000 W
Logic (RQL, $I_c = 25 \mu A$, 8.3 GHz) • processor cores	0.18 W • 40,200	1.8 W • 402,000	18 W • 4,020,000	180 W • 40,200,000
Memory (1 B/FLOP, JMRRAM) • quantity (1 B/FLOPS)	0.46 W 1 PB	4.6 W 10 PB	46 W 100 PB	460 W 1,000 PB
Interconnects (VCSELs @ 40 K)	0.1 W	1 W	10 W	100 W
Other (structure, radiation heat leaks)	0.76 W	2.6 W	26 W	260 W
Total	1.5 W	10 W	100 W	1,000 W
• Computation efficiency (goal: $\geq 5 \times 10^{11}$ FLOPS/W)	0.7×10^{11} FLOPS/W	2.5×10^{11} FLOPS/W	5×10^{11} FLOPS/W	5×10^{11} FLOPS/W



Conclusions:

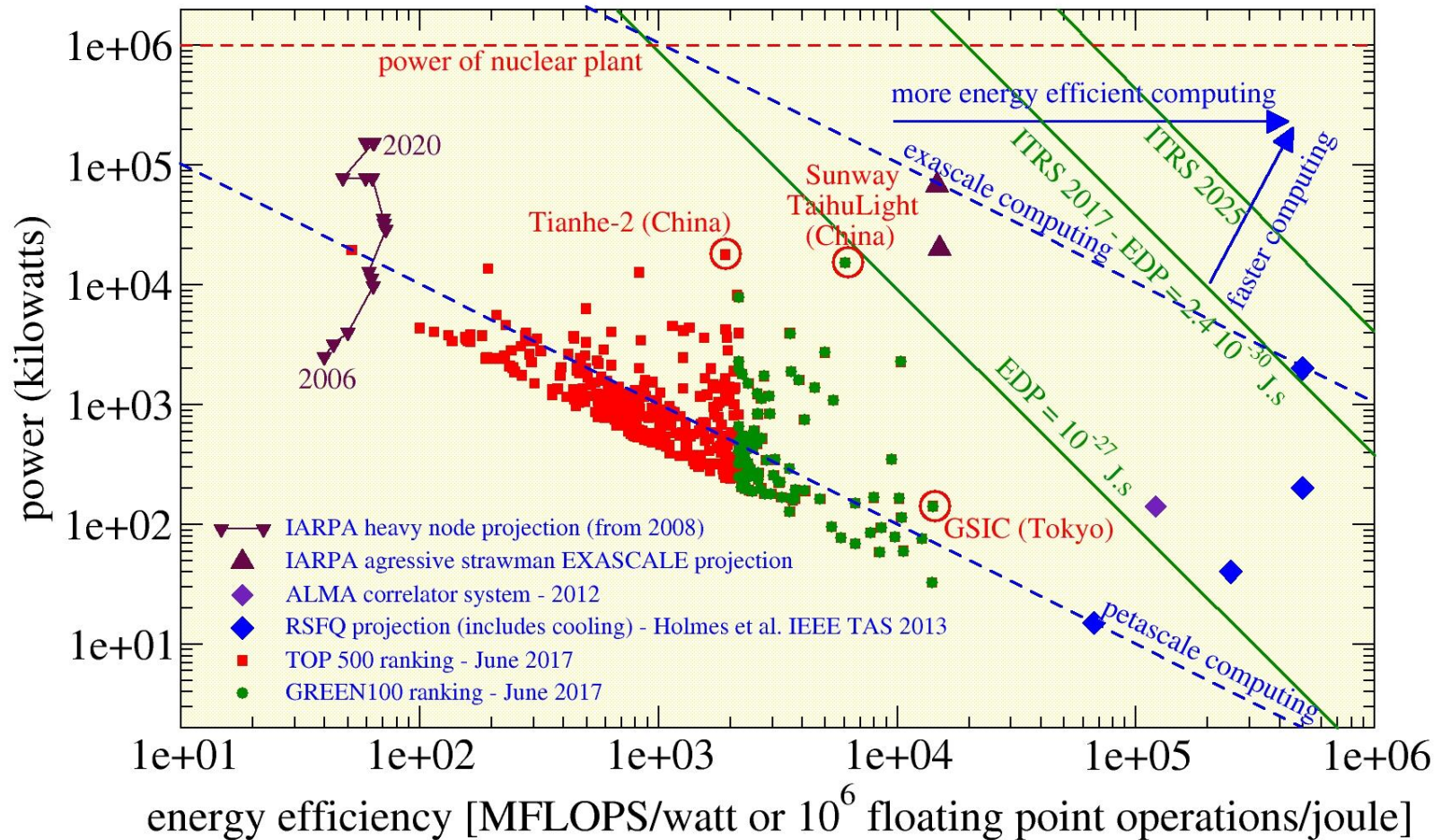
- Energy-efficient superconducting computers are possible
- Priorities:
Memory → **Logic** → **System** → **Interconnects**

Source: Scott Holmes - Superconducting SFQ VLSI Workshop (SSV 2013) - November 2013

Comparison of high-computing technologies

Performance and power of high-end computing

September 2017 - total power consumption : 0.65 GW



Cryogeny : Stirling cryocoolers

Long experience gained with tactical mini-coolers for infrared detection

Sliding rotating or linear pressure oscillator + mechanically or pneumatically driven cold expander: from a few 1000h up to a few 10.000h MTBF, $\frac{1}{4}$ to 2W @ 80K



Source: Alain Ravex -Absolut Systems

Superconducting digital electronics integrated systems



Complete cryocooled digital-RF receiver system prototype, assembled in a standard 1.8-meter tall 0.5-meter wide equipment rack.

Using the modular packaging approach, the system can currently host variety of chips.

The system includes a two-stage 4-K Gifford-McMahon cryocooler manufactured by Sumitomo, two sets of interface amplifiers for connecting chip outputs to an FPGA board (placed behind the vacuum enclosure, on the metal tray) for further digital processing and computer interface. The system also includes a current source and a temperature controller.

Courtesy of Deep Gupta - HYPRES

Commercially available 4K cryorefrigerators

Commercial 4K cryorefrigerators implementations :

- lubricated screw type helium compressors with oil

injection at suction and oil removal system at exhaust

- aluminum plate counter flow heat exchangers
- gas bearings frictionless cold expanders

- Typical characteristics:

- automatic operation
- efficiency: 20 - 25% of Carnot
- turbines reliability: > 100.000h MTBF
- cooling capacity/electrical input:
from 100W/50kW up to 1 kW/250kW @ 4K

Source: Alain Ravex - Absolut Systems



Lessons learnt for large cooling power from LTS high energy physics projects (i.e. CERN/LHC)

64 compressors ($39\text{MW}_{\text{elec}}$)
74 cold expansion turbines
28 cold compressors
1200 current leads
1800 sc magnets

95% cryogenic system availability

30% Carnot efficiency

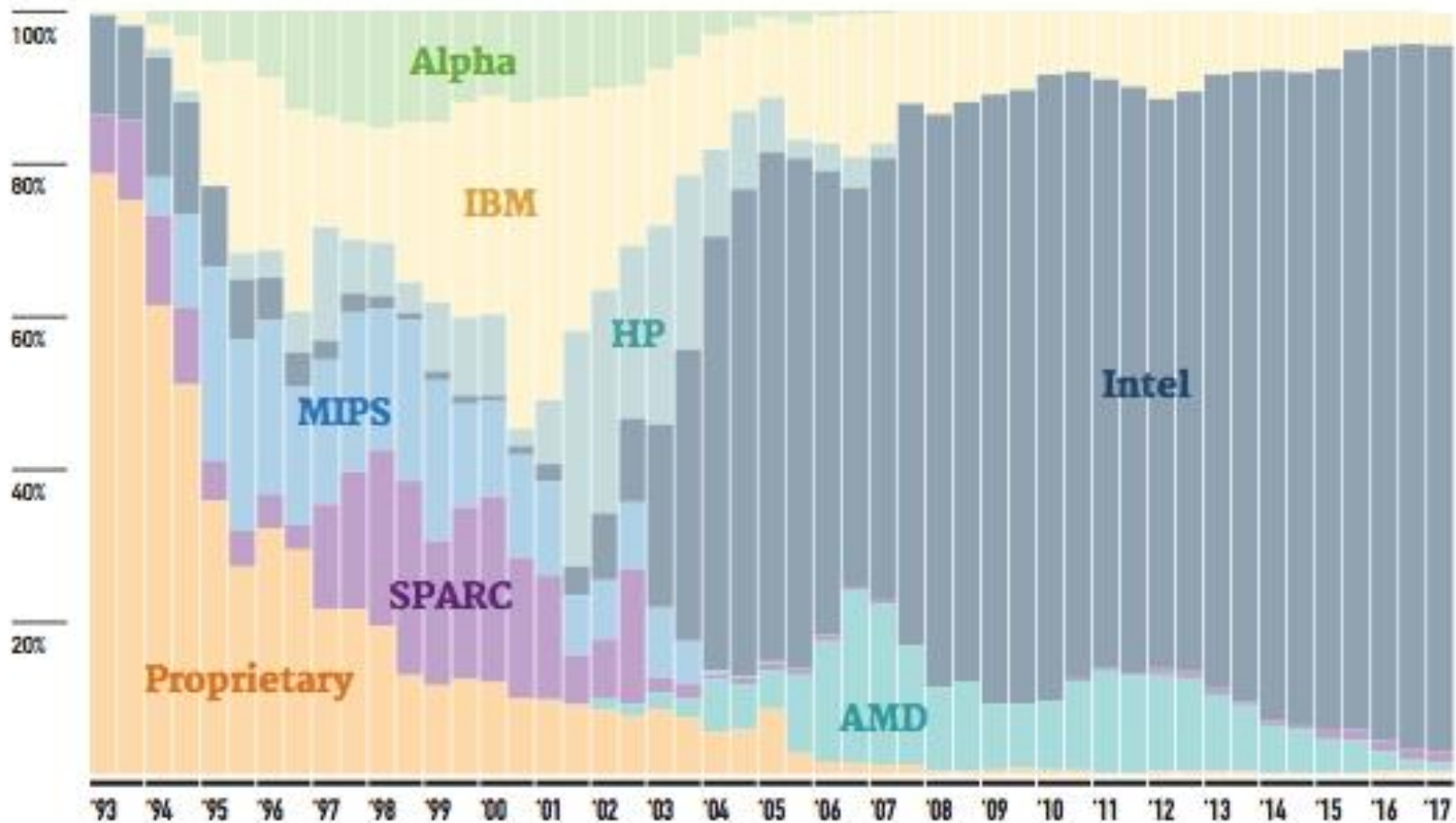
Industrial type operation with high efficiency, reliability and availability demonstrated for large cooling capacity 1.8K cryorefrigerators

Source: Alain Ravex - Absolut Systems



Next decade : how to proceed?

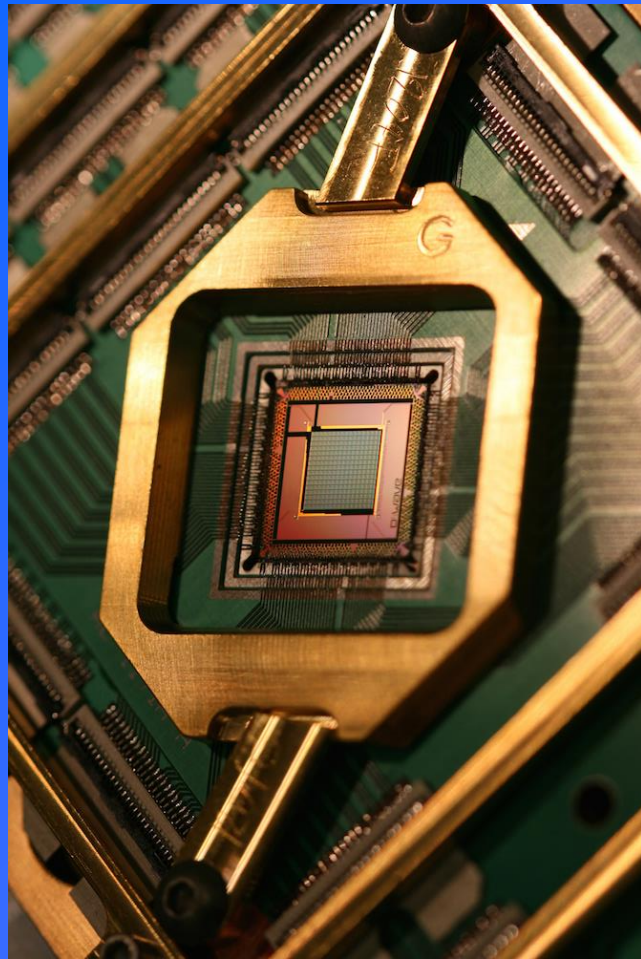
CHIP TECHNOLOGY



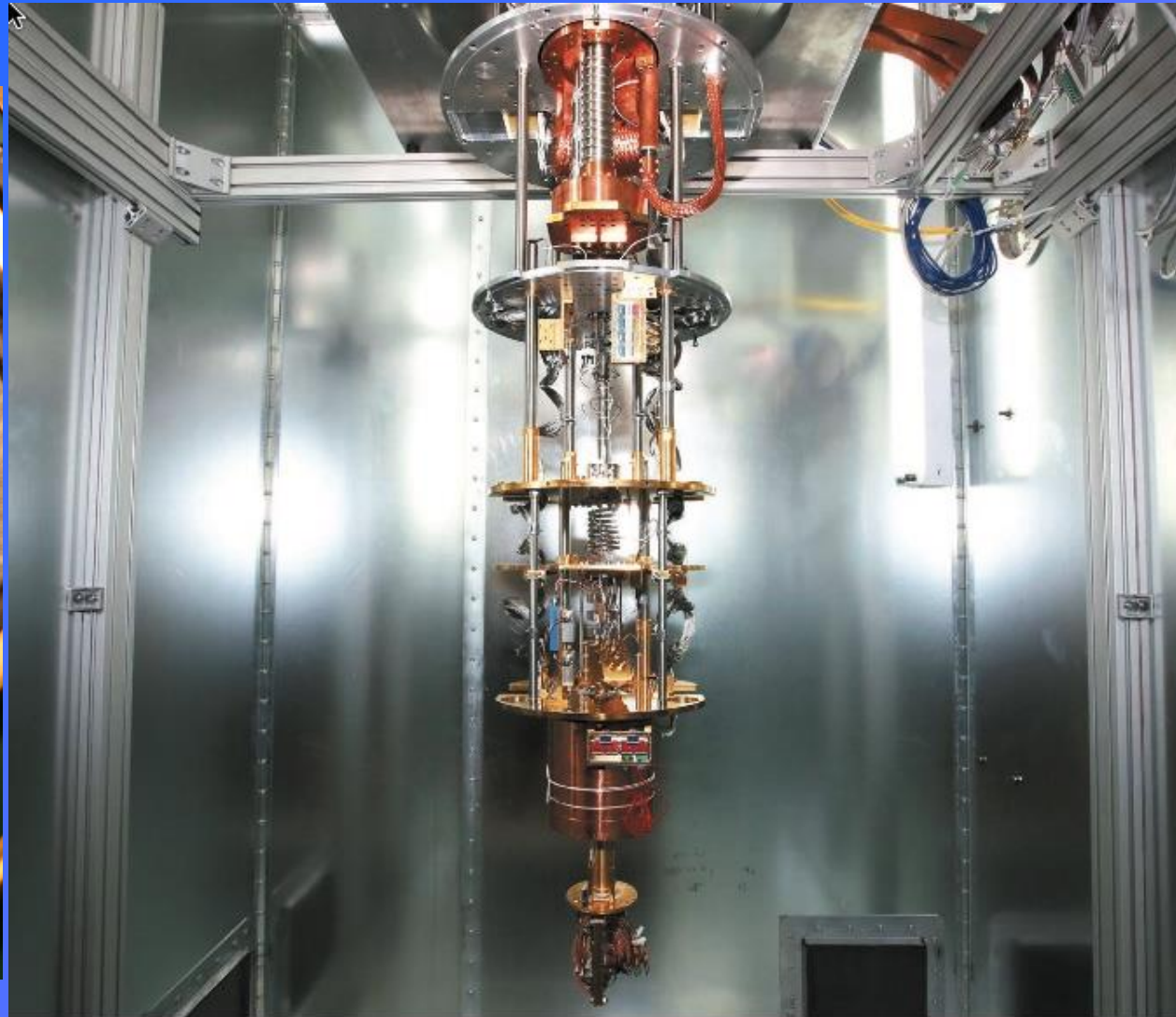
hold the promise to solve computational problems beyond current supercomputers.

Source : <https://ec.europa.eu/digital-single-market/en/high-performance-computing>

Will next decade be quantum?



Source : D-Wave



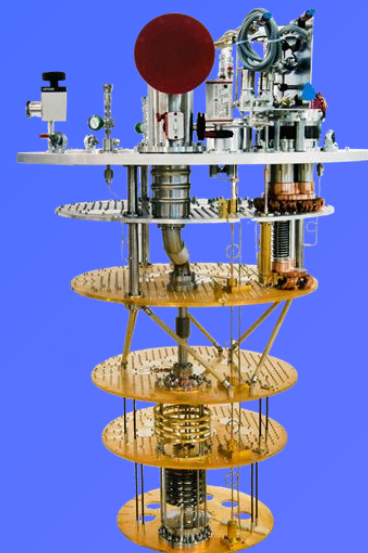
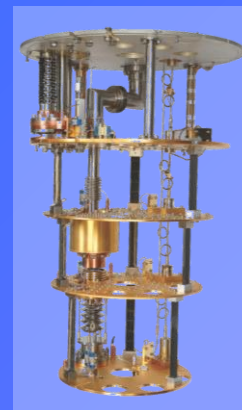
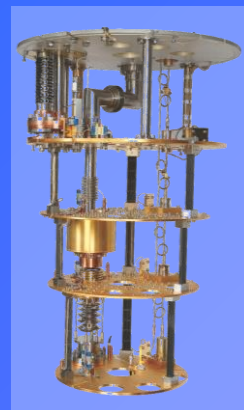
D-Wave's latest processor has 2,000 qubits — far surpassing the capacity of previous models.

Source : E. Gibney, "Quantum computer gets design upgrade," Nature, January 2017

Switching energies - speed


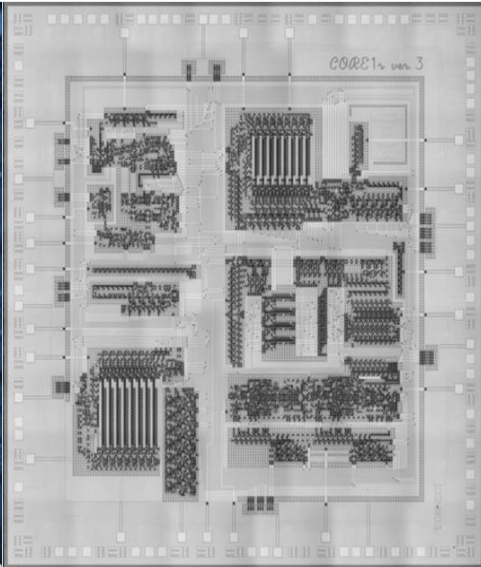
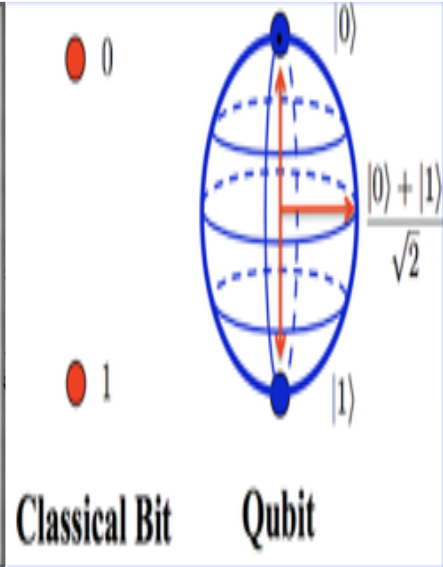
Technology	CMOS	superconducting digital electronics	quantum computing
temperature	300	4	0.04
thermal energy (J)	$4 \cdot 10^{-21}$	$5.5 \cdot 10^{-23}$	$5.5 \cdot 10^{-25}$
switching energy (J)	10^{-16}	10^{-19}	-
"thermal" frequency	6 THz	83 GHz	0.8 GHz

Cryogen free ultra low temperature coolers



Product name	lo	Triton 300	Triton 500	XL1000
Base Temperature	40 mK	10 mK	10 mK	3.3mK
cooling power at 10 mK				5 μ W
cooling power at 20 mK		3 μ W	12 μ W	25 μ W
cooling power at 100 mK	25 μ W	300 μ W	500 μ W	1000 μ W
Cooling power at 1K	200 mW	10 mW	10 mW	10 mW
Cooling power at 4K	1 W	1 W	1 W	3 W
Base plate for base temp (mm)	150	290	290	430
PTR type/number	1 W	1.5 W	1.5 W	2x1.5 W

From CMOS to quantum computers

	<i>CMOS processing</i>	<i>superconducting processing</i>	<i>quantum processing</i>
<i>technology</i>			
<i>energy efficiency</i>	10^4 MFLOPS/W	$5 \cdot 10^5$ MFLOPS/W (includes cooling)	??
<i>speed</i>	2.4 GHz	100 GHz	problem-dependent
<i>gates per cm²</i>	$2.64 \cdot 10^9$	$0.8 \cdot 10^6$	$2 \cdot 10^3$
<i>temperature of operation</i>	300 K	4 K	0.01 K

Conclusion and prospectives

- Superconducting digital electronics has achieved some major breakthroughs during the last decade that enables the fabrication of superconducting microprocessors today (Core e2, Core e4) through energy-efficient biasing techniques and higher integration.
- Some challenges are still ahead :
 - memory issue
 - further integration required by down-sizing gates : higher square inductances (NbN, thicker films), narrower line widths (down to what values?), 3D integration.
 - design tools for complex circuits need be developed
- Some prospectives :
 - other materials need more investigation, either for JJ or/and for interconnects : NbN, MgB₂
 - a small increase of temperature of operation would be of great help for energy budget (4K → 10K → 20K)
 - exascale objective may not be ambitious enough regarding semiconductors advances : a larger cryogenic system is more energy-efficient.
 - Superconducting digital electronics is a natural interface between room-temperature semiconductors electronics and quantum computing systems