Software and Computing Domain: WP2.6.2 Computing Hardware Architecture

Duncan Hall, Chris Broekema

WP2 Meeting

2010 October 29

Outline

WP2.6.2 Hardware options for SKA computing

- Calibration and imaging requirements for SKA1
- Forecasts of COTS hardware capabilities
- Forecasts of COTS power requirements
- Data input – output challenges

Outline

WP2.6.2 Hardware options for SKA computing

- Calibration and imaging requirements for SKA1
- Forecasts of COTS hardware capabilities
- Forecasts of COTS power requirements
- Data input – output challenges

WP2.6.2 Computing hardware architecture development: requirements for calibration and imaging

Contributions to date; Current Status	Challenges to be addressed
•A spreadsheet model identifying key cost driver parameters and likely pessimistic-optimistic ranges has been published on the S&C Domain wiki	•The required number of floating point operations per u-v sample ("flops per float") will likely range from 100,000 to 400,000+ •This leads to requirements for hundreds of petaflop/s to ~1 exaflop/s for SKA1 as defined in Memo 125 •The "flops" metric is only one measure of HPC performance; disk input-output data rates, cache memory access speeds and sizes, and hardware reliability can be just as important

CPG Memo 3 (2009-11-6) confirms requirements for extreme scale computing:

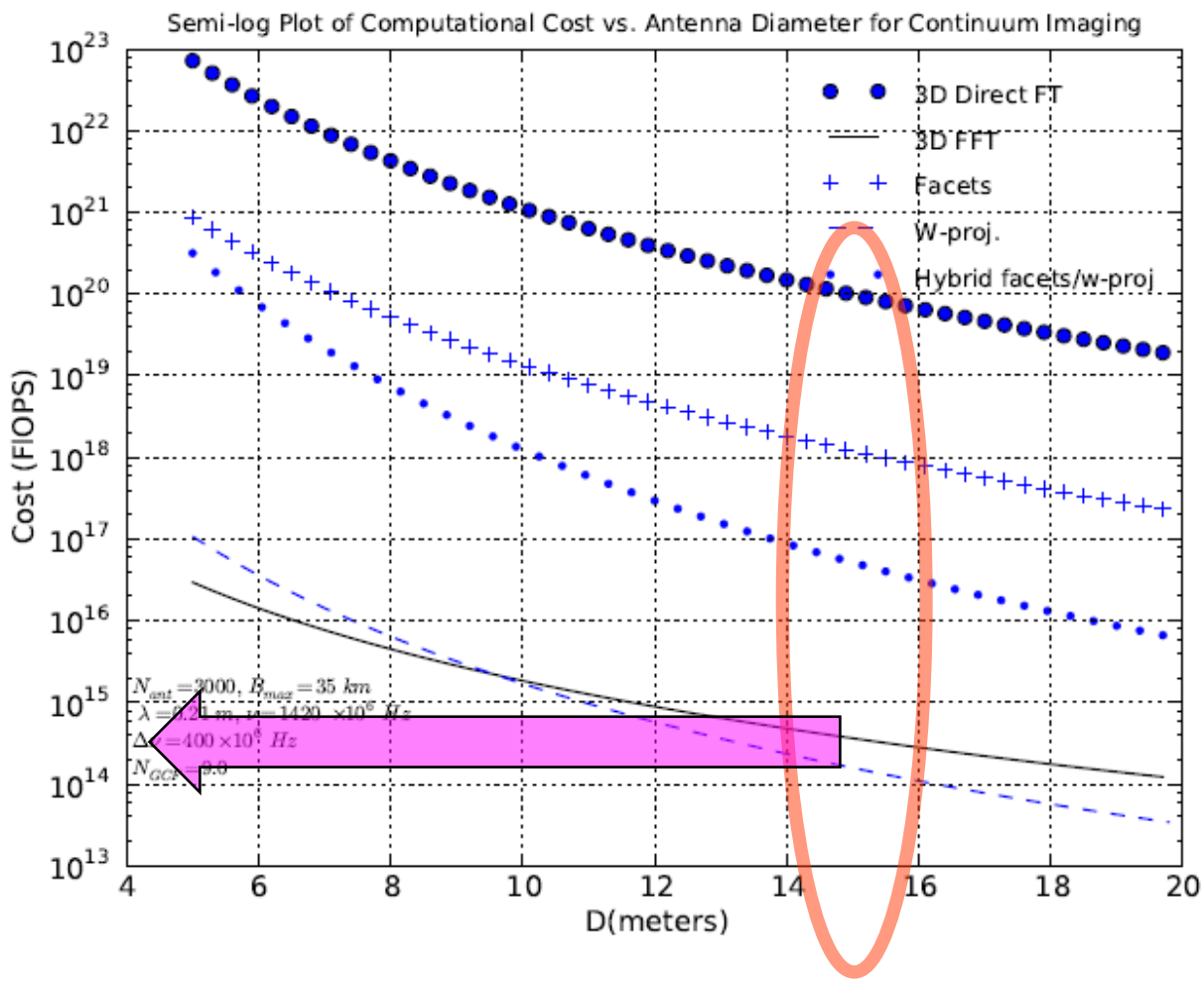
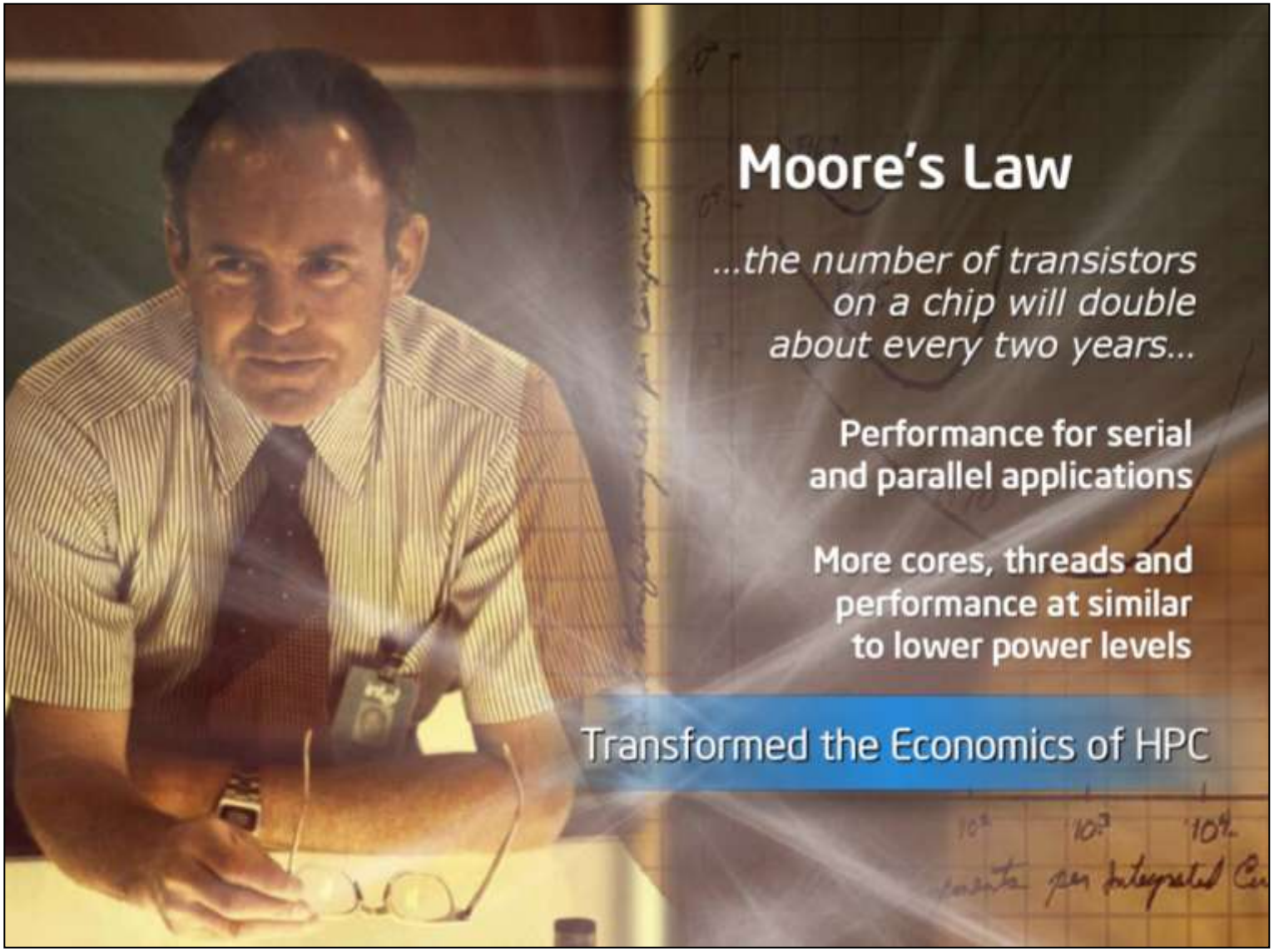


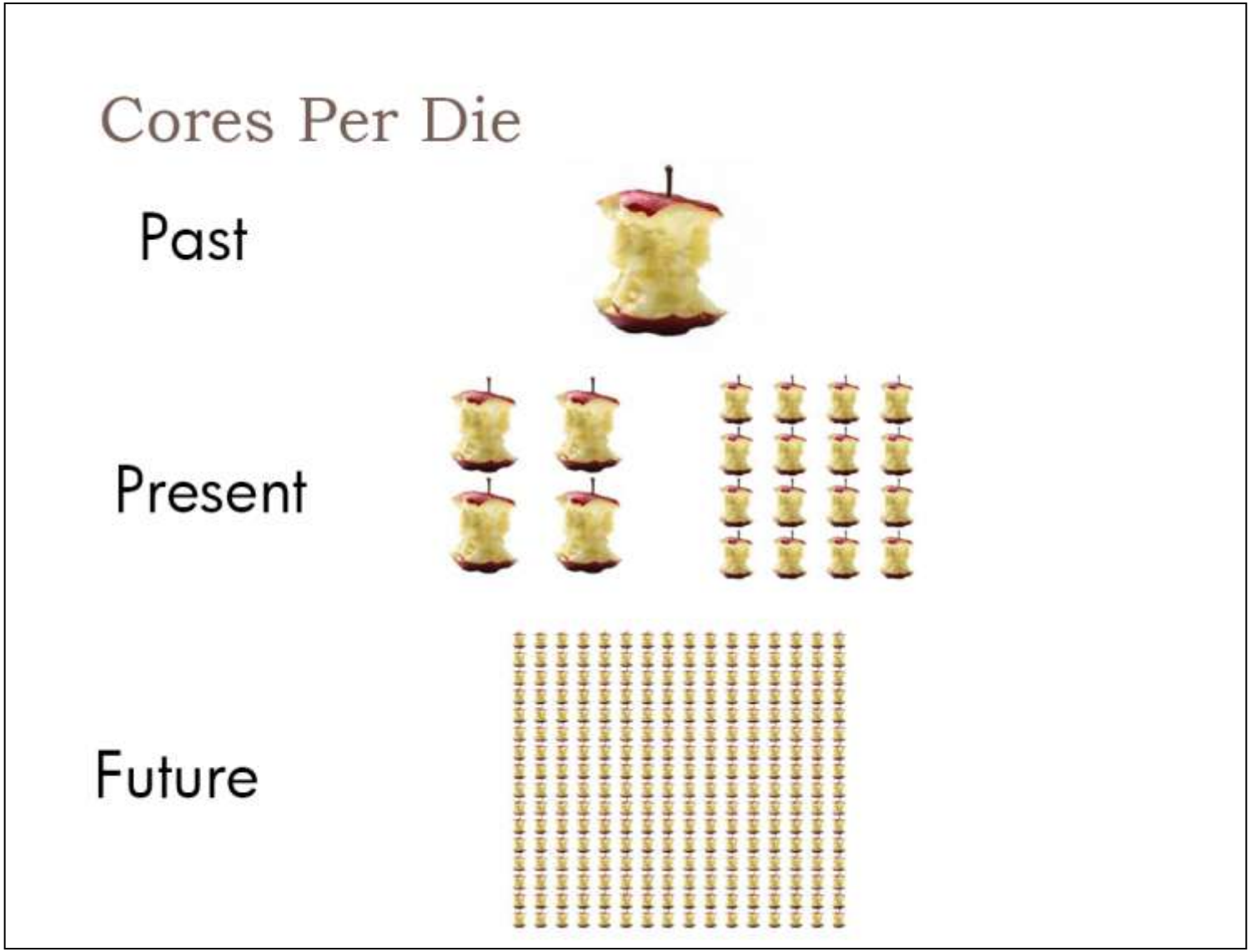
Figure 1: Semi-log y plots of computational costs (without consideration of deconvolution and parallel computing efficiency η) vs. antenna diameter D for continuum imaging for the 3-D direct FT, 3-D FFT, facets, w-projection, and hybrid facets/w projection imaging algorithms.

Intel asserts that Moore's 1965 observation will continue to hold to ~2020 ...



From a presentation delivered by Intel's Kirk Skaugen at ISC'10, 2010 May 31

... but the advent of multiple cores per chip has major implications for software at extreme scale ...



From a presentation delivered by HP Labs' Norman Jouppi at SC'09, 2009 November 15

... and Amdahl's laws override Moore's law:

Amdahl's Laws

Gene Amdahl (1965): **Laws for a balanced system**

i. Parallelism: max speedup is S/(S+P)

ii. One bit of IO/sec per instruction/sec (BW)

iii. One byte of memory per one instruction/sec (MEM)

Modern multi-core systems move farther away from Amdahl's Laws (Bell, Gray and Szalay 2006)

From a presentation delivered by Alex Szalay at XLDB4, 2010 October 7

WP2.6.2 Computing hardware architecture development: requirements for calibration and imaging

Challenges to be addressed	Work to be done; Milestones; Risks
•The required number of floating point operations per u-v sample ("flops per float") will likely range from 100,000 to 400,000+ •This leads to requirements for hundreds of petaflop/s to ~1 exaflop/s for SKA1 as defined in Memo 125 •The "flops" metric is only one measure of HPC performance; disk input-output data rates, cache memory access speeds and sizes, and hardware reliability can be just as important	•Opportunities to improve performances over that of current codes must be explored – in WP2.6.3 work •Opportunities to shift computing load from general purpose von Neumann architecture to special purpose architectures – using for example hardware accelerators – must be explored as a follow on from WP2.6.3 work

Outline

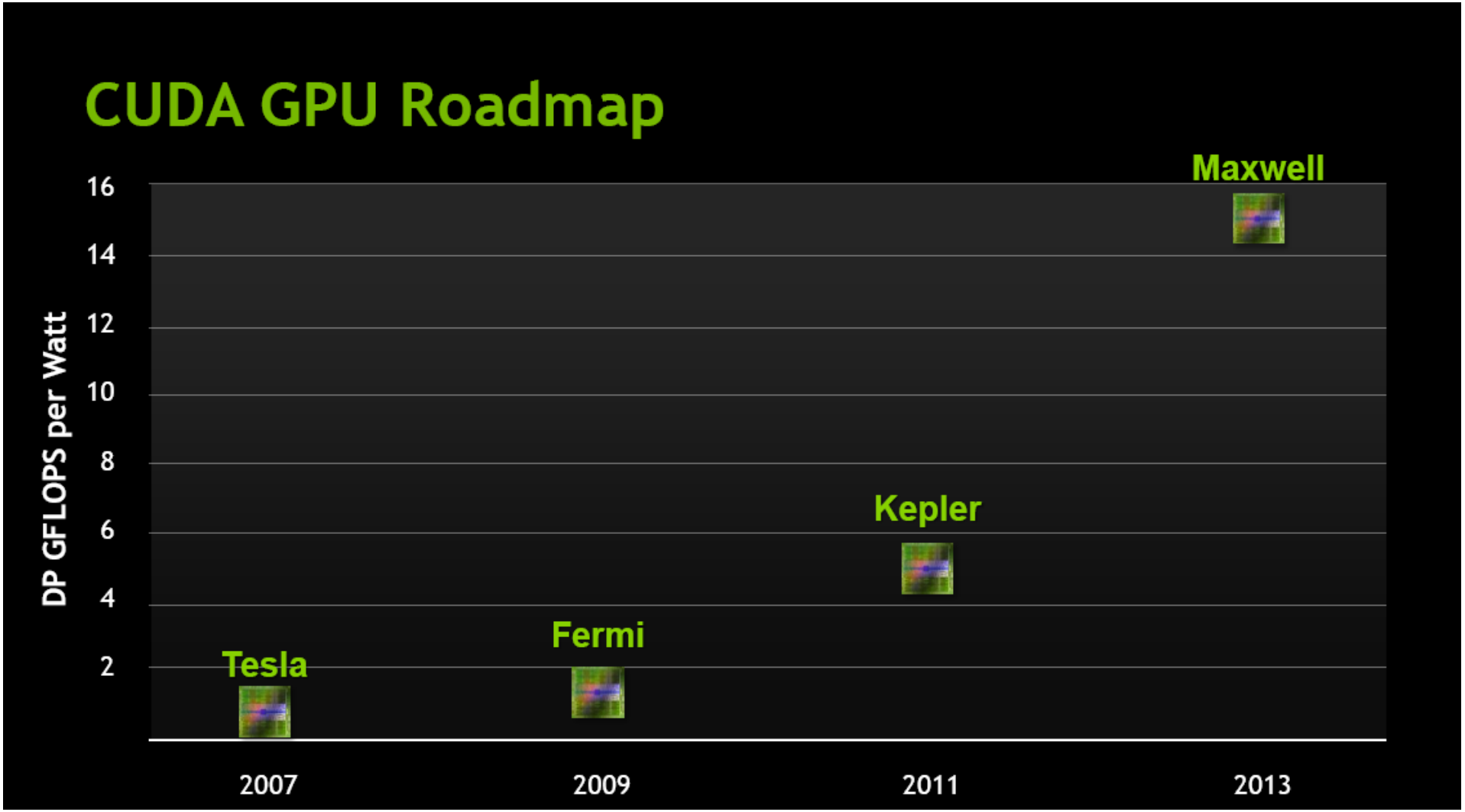
WP2.6.2 Hardware options for SKA computing

- Calibration and imaging requirements for SKA1
- Forecasts of COTS hardware capabilities
- Forecasts of COTS power requirements
- Data input – output challenges

WP2.6.2 Forecasts of COTS hardware capabilities and power requirements

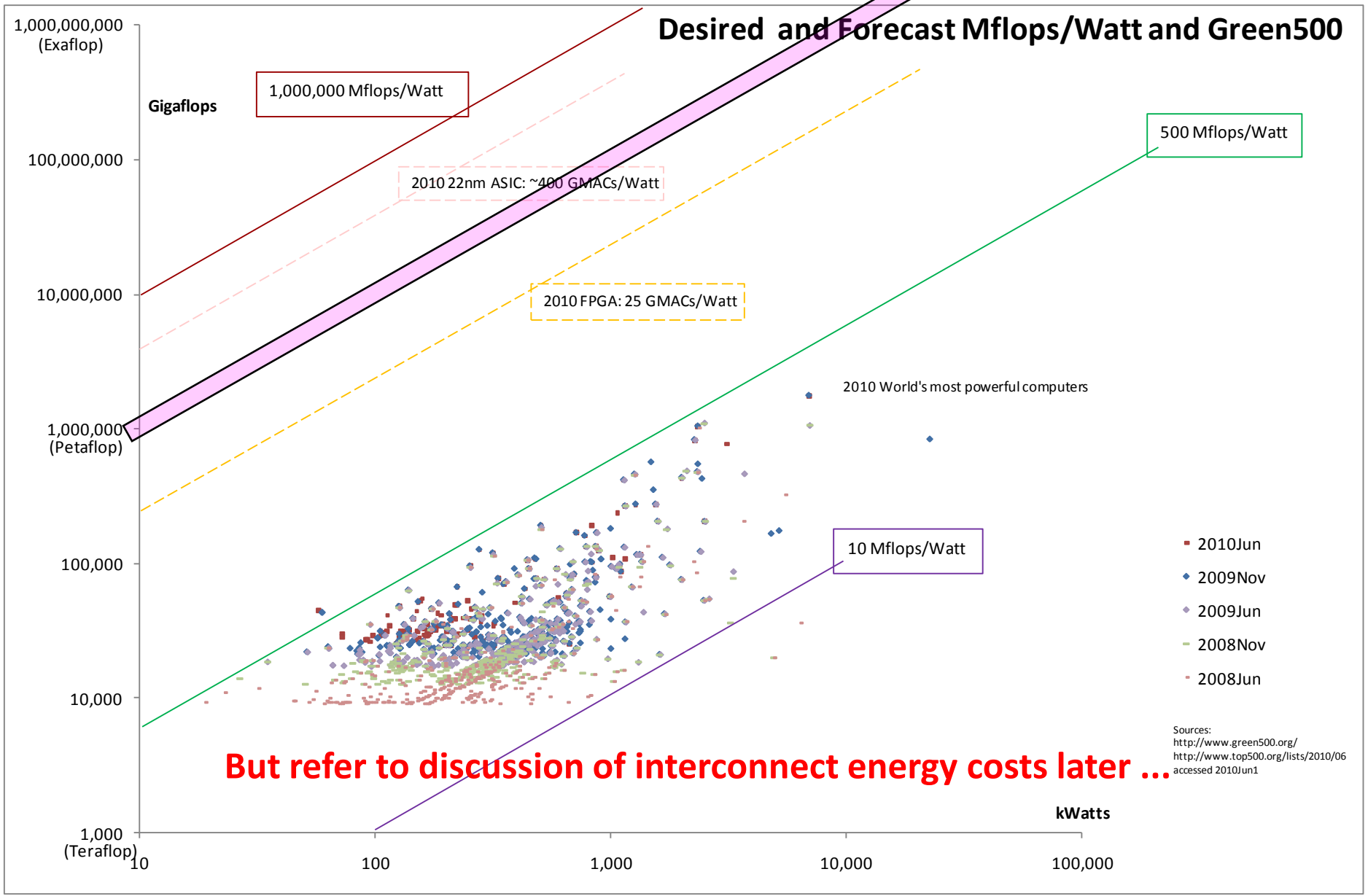
Contributions to date; Current Status	Challenges to be addressed
•Recent published research on Commercial Off The Shelf (COTS) hardware capabilities and power requirements has been posted on the S&C Domain wiki •Forecasts of performance from vendors of compute elements (e.g. Intel and NVIDIA) have been posted on the S&C Domain wiki	•Other than essentially one-off “icon” installations, COTS High Performance Computers (HPCs) with capacities of ~1 exaflop/s are not likely to be commercially available until ~2020 •The typical cost of high end HPC hardware is ~€100 million; other infrastructure such as persistent storage is additional •The stretch target power consumption for exaflop/s class HPCs given to US vendors is ~20 MW for the HPC alone

Forecast: ~135 watt Maxwell chip to deliver ~2 DP Tflop/s in 2013
SP ≈ 5 x DP ⇒ ~100 SP Gflop/s per watt [= 10pJ/flop]



From a presentation delivered at 2010 September 20-23 GTC conference by Jen-HsunHuang of NVIDIA

It is not impossible that the equivalent of GPU-powered boards operating at 100+ SP Gflop/s per watt could deliver 1 EF for ~10 MW



But refer to discussion of interconnect energy costs later ...

WP2.6.2 Forecasts of COTS hardware capabilities and power requirements

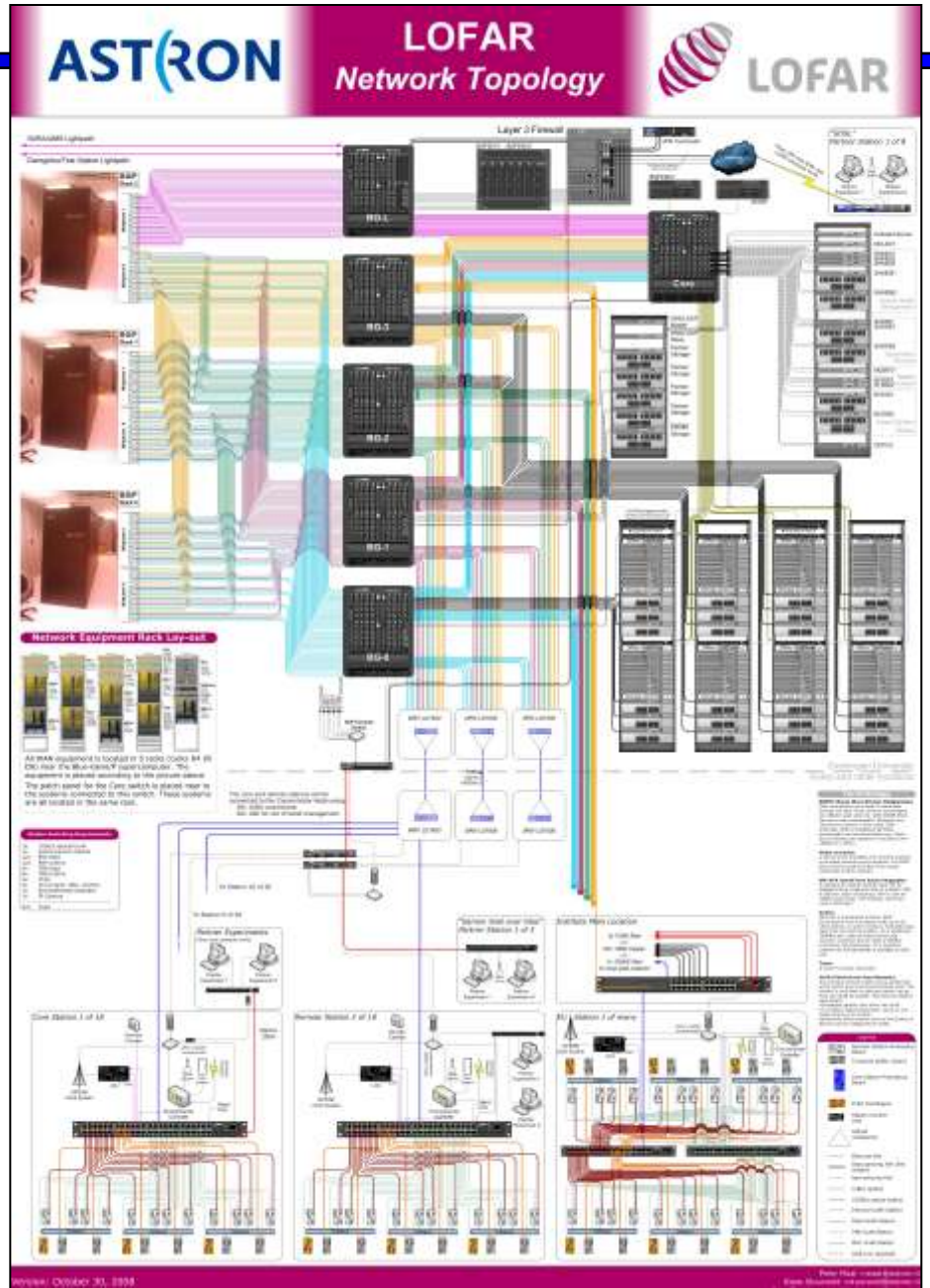
Challenges to be addressed	Work to be done; Milestones; Risks
•Other than essentially one-off "icon" installations, COTS High Performance Computers (HPCs) with capacities of ~1 exaflop/s are not likely to be commercially available until ~2020 •The typical cost of high end HPC hardware is ~€100 million; other infrastructure such as persistent storage is additional •The stretch target power consumption for exaflop/s class HPCs given to US vendors is ~20 MW for the HPC alone	•Characterisation of all sources of power consumption – including interconnection, power conditioning, persistent storage and environmental control – is required in order to estimate the total power costs •Significant work is required to assess the feasibility of HPC using heterogeneous architectures, e.g. learn from LOFAR, ASKAP and Single Digital Backend •Interim results of work to be delivered before CoDR in September 2011 •Little downside risk: work is research

Outline

WP2.6.2 Hardware options for SKA computing

- Calibration and imaging requirements for SKA1
- Forecasts of COTS hardware capabilities
- Forecasts of COTS power requirements
- Data input – output challenges

Network and interconnection for LOFAR is “non-trivial”



WP2.6.2 Data input-output challenges

Contributions to date; Current Status	Challenges to be addressed

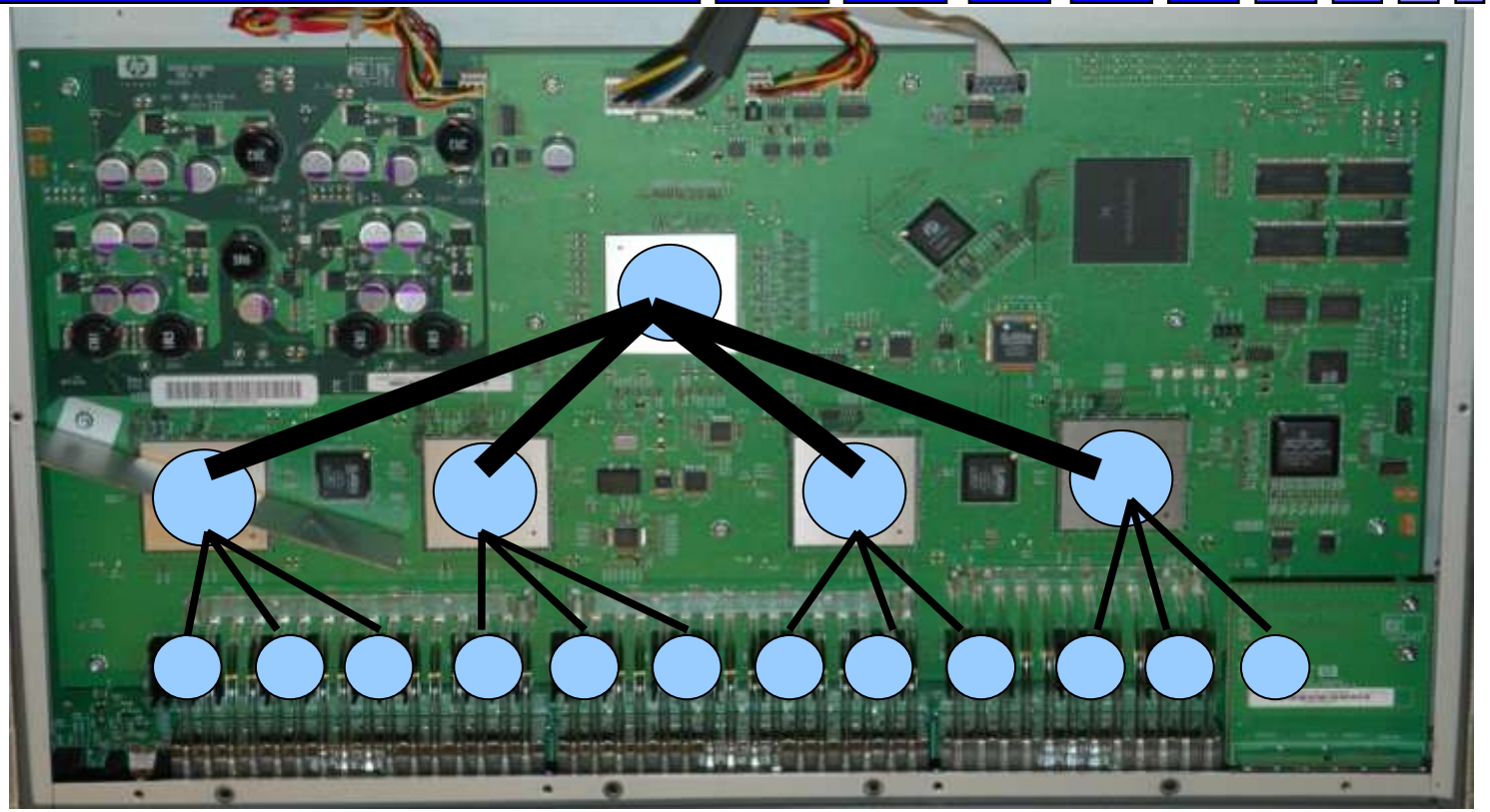
•ASTRON experience: for LOFAR data streaming, the default IBM Blue Gene P input I/O configuration was more than 50% too slow
•Various optimisations were required to improve performance
•Need to consider I/O on all levels:
  •System design, operating system, communications stacks, application
•Current software not optimised for throughput:
  •Mostly optimised for connections and stability
  •Much of the required work needs to be done in-house
  •At least some expert knowledge is required, e.g. Linux kernels on I/O nodes
	•Cost of I/O is non-linear with scale:
	  •<u>Number of active switch chips</u>: scales faster than port count because of fat-tree configuration; doesn't scale beyond a few thousand ports due to limited port count per component with much worse scaling from there on
	  •<u>Procurement cost</u>: added costs of interconnect components; additional complexity; and high bandwidth COTS cards are just not available beyond a few thousand ports
	  •<u>Energy</u>: line energy cost is ~linear with distance, but need to add costs of noise handling algorithms and interconnect active components

Scaling issues with standard network port cards (1/2)

A 48 port GbE switch, representative of a hardware-efficient universal fat-tree network design:

- Loads of chips needed already
- Non-blocking performance not guaranteed

Scaling issues with standard network port cards (2/2)

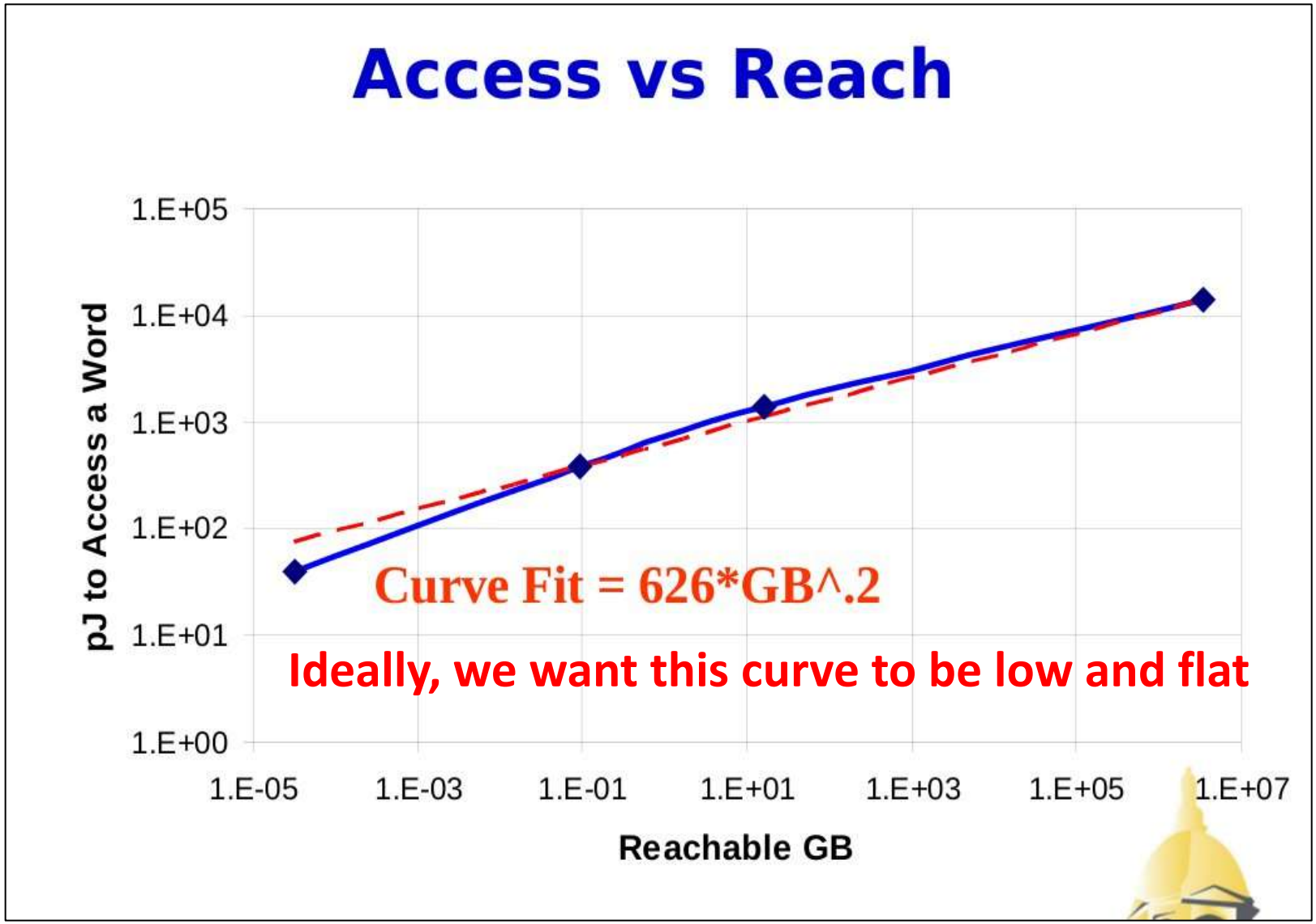


This design will not scale indefinitely.

For N ports per switch chip:

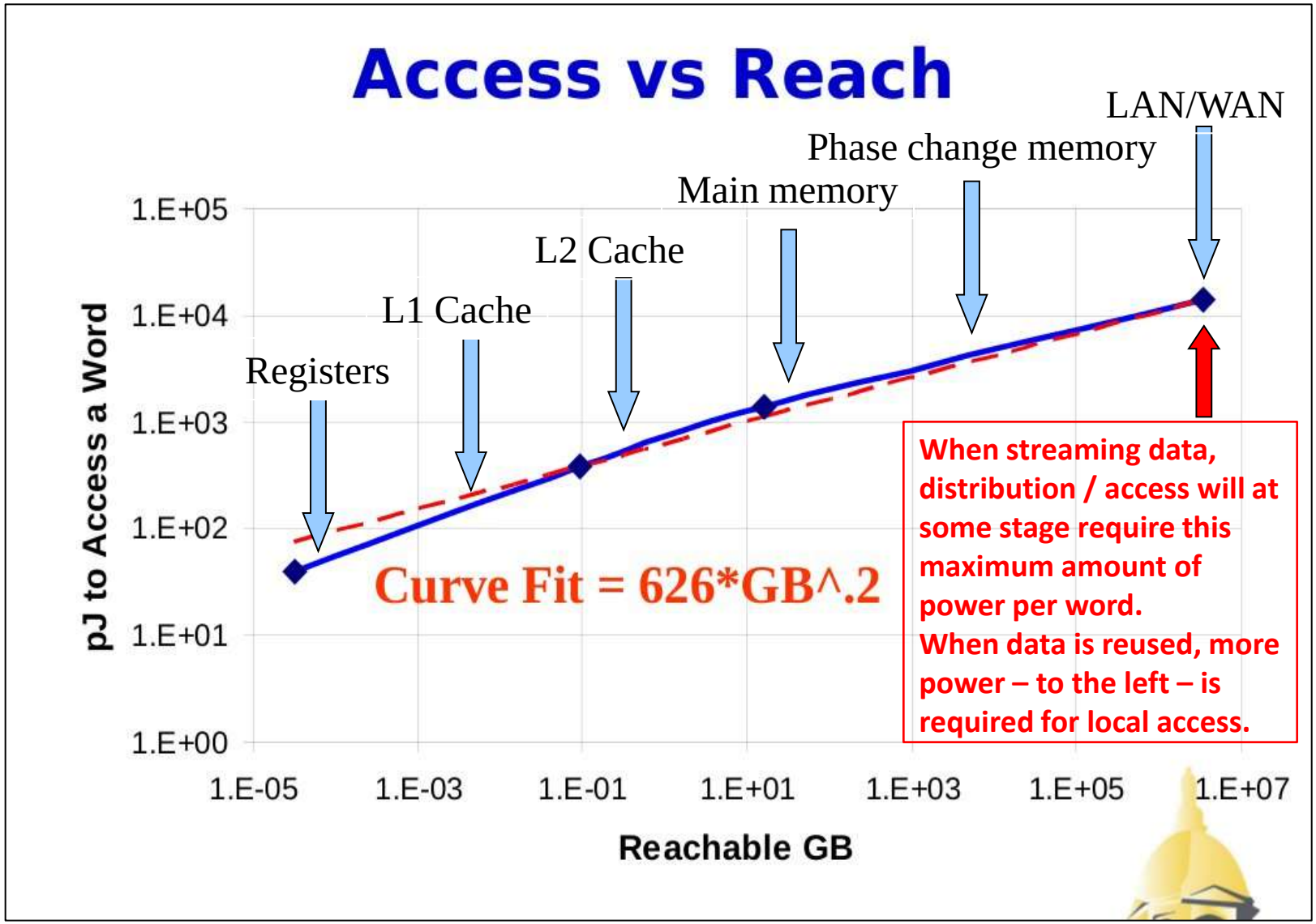
Max port count = $N^2 / 2$

when using a single type of switch chip; this uses three types



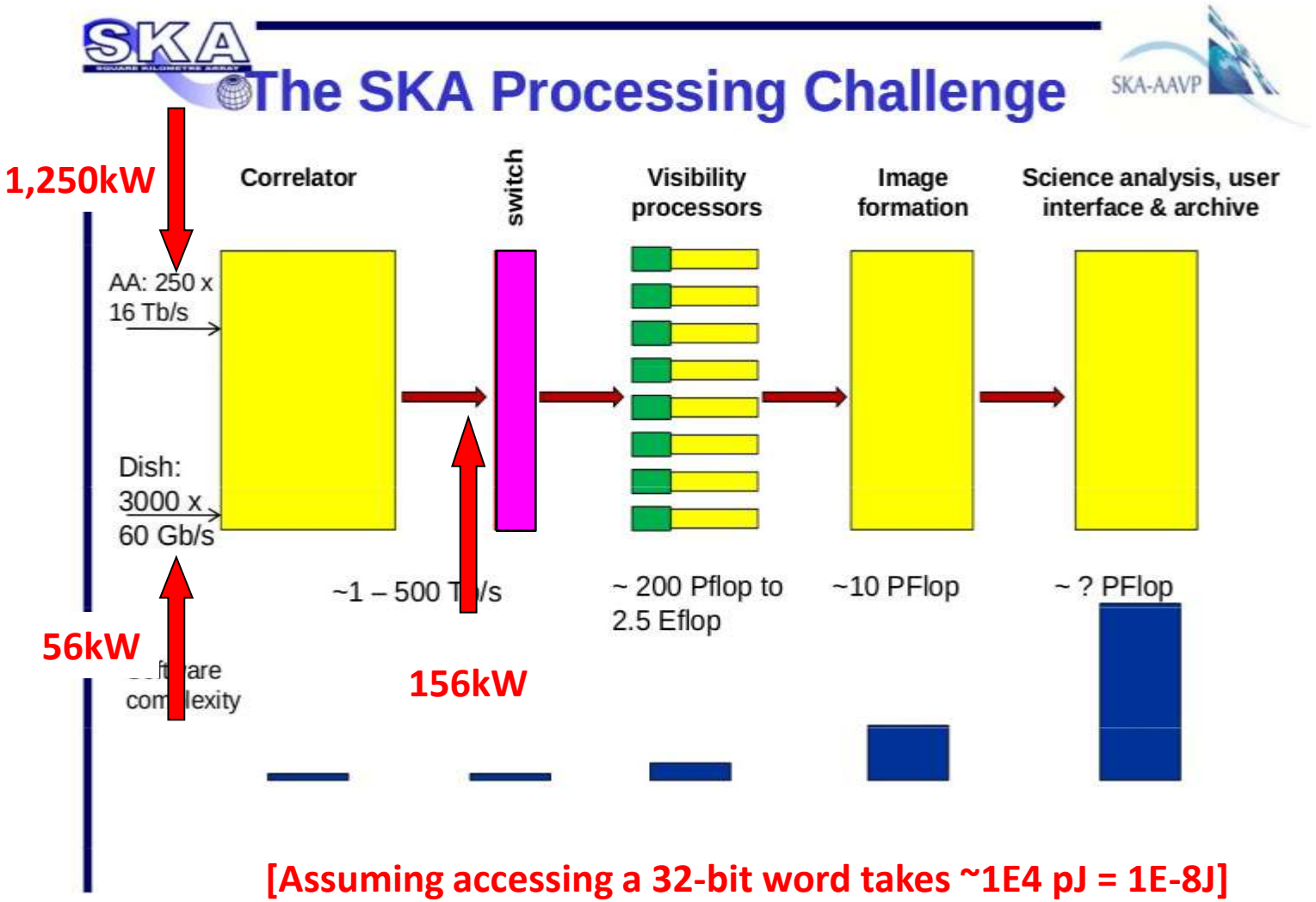
Source: Energy at ExaFlops, Peter M. Kogge, SC09 Exa Panel

WP2.6.2 energy cost input / output



Source: Energy at ExaFlops, Peter M. Kogge, SC09 Exa Panel

WP2.6.2 energy cost for input / output: interconnection at top level only



WP2.6.2 energy cost for input / output: summary of challenges and work to be done

Moving data costs a tremendous amount of energy

- Hundreds of kilowatts to megawatts for SKA
- Moving data over hundreds or thousands of km will cost even more energy

Switching costs a lot of energy too

- A non-blocking thousands-of-ports switch will consume power at a staggering rate

This will probably not scale well

- N-dimensional torus may be feasible; but is not currently COTS

In addition, getting data into various machines is challenging

- see LOFAR case

Limiting the port count can significantly reduce cost

- By avoiding low efficiency network topologies
- Should investigate design of an over-subscribed network

Finally, need to reduce data flows to keep SKA affordable

- Otherwise moving bits will end up dominating the power budget

Key messages:

- **Sustained COTS exascale is coming: ~2020**
  - US and other government-funded initiatives for e.g. energy and climate modelling and related research
  - Substantial change in hardware architectures will drive change in software: e.g. multiple threading across millions of cores (IESP)
- **But, SKA1's requirements will push the 2020 envelope of COTS hardware capabilities**
- **Purpose built hardware solutions for SKA1 will also be subject to the challenges of computation at exascale:**
  - Amdahl's laws
  - Energy for interconnection
  - Reliability