Extreme Computing: Challenges, Constraints and Opportunities

Professor Anne E Trefethen

Oxford e-Research Centre

Oxford University, UK

Outline

- Trends and roadmaps for extreme computing
- Co-design vehicles – the Square Kilometre Array
- Achieving realist energy efficiency
- Conclusions

Trends in Extreme Computing

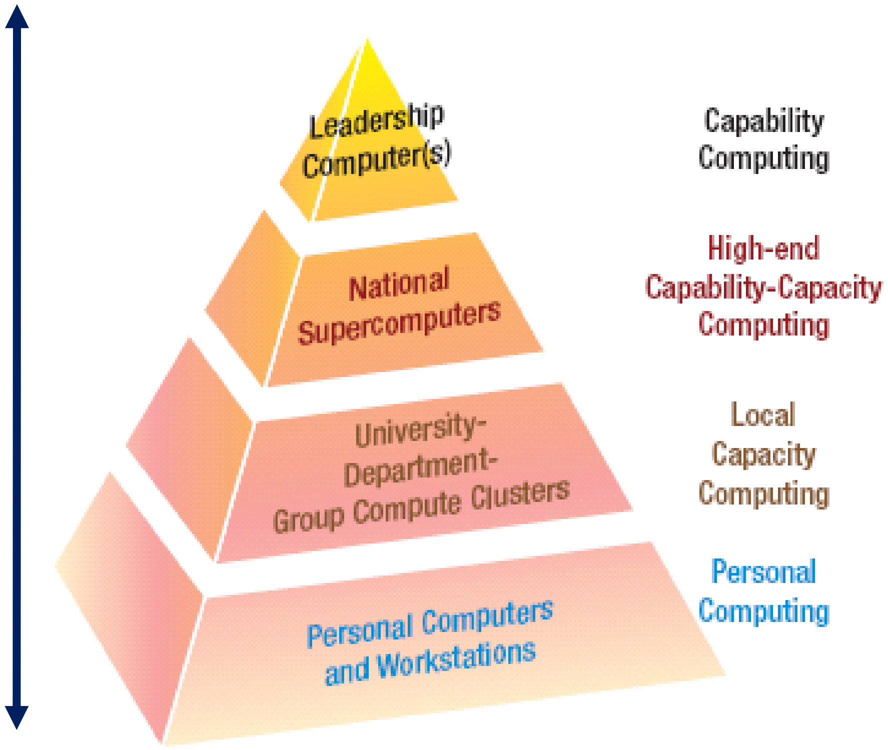
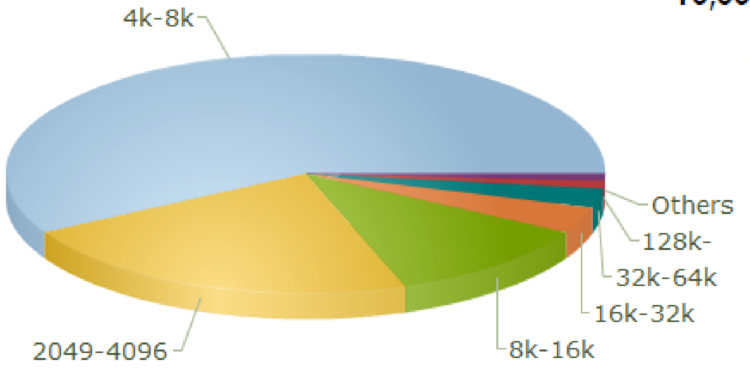


Figure 1. The spectrum of computing resources.

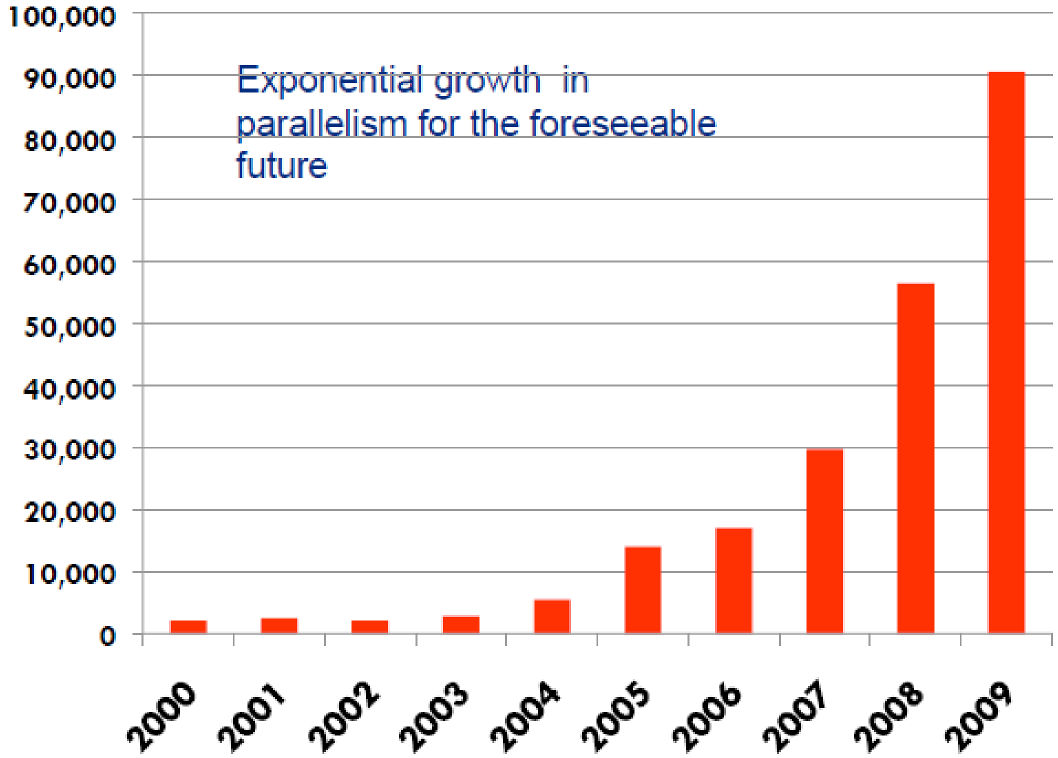
- Heterogeneity from desktop to high-end systems creating complexity in efficient application development
- Multi-core beginning to dominate
- Grids, clouds and hpc show signs of converging

Top500 processor numbers

Number of Processors / Systems
June 2010



Top20 of the Top500



Exponential growth in parallelism for the foreseeable future

(Courtesy Jack Dongarra)

Roadmaps for Extreme computing

- UK HPC/NA Roadmap
  - – http://www.oerc.ox.ac.uk/research/hpc-na
- European Exascale Software Initiative (EESI)
- International Exascale Software Project (IESP)
  - – http://www.exascale.org

Aims of the Roadmapping Activity

Survey a range of applications and users to understand:

- The role and limits of a common algorithmic base
- How this common algorithmic base is currently delivered and how should it be delivered in the future
- What are the current requirements and limitations of the applications, and how these should be expanded
- What are the “road-blocks” that limit the scope of the future exploitation of these applications.
- A better comprehension of the “knowledge gap” between algorithmic developments and scientific deployment
- How significant computing language as well as other “practical” issues weigh in the delivery of algorithmic content

Activity to date

Community Consultation

- Workshop 1: Oxford, Nov 2008
  - Applications focus
- Workshop 2: Manchester, Dec 2008
  - Algorithms/NA focus
- Workshop 3: London, Jan 2009
  - Review of Roadmap V1, further user & industry perspectives

Background work

- considering DOE/DARPA/NSF workshops
- Discussions with applications outside of workshops

Vision for the Roadmap

The Grand Challenge is to provide

- software that application developers can reuse in the form of high-quality, high-performance, sustained software libraries and modules
- a community that allows communication of interdisciplinary knowledge, and the development of appropriate skills.

The Roadmap has identified main five themes for action

Theme 1: Cultural Issues

There is a need to

- Identify potential community players across application domains, numerical analysis and computer science
- Develop models of community sharing of algorithms, software and ideas
- Provide community activities, workshops, training, virtual meeting spaces.
- Engage internationally

No one community can address all the issues alone – we need international, interdisciplinary teams

Theme 2: Applications and Algorithms

There is a need to:

- Identify exemplar applications to develop baseline models for communication and benchmarking
- Develop a map of algorithms across application domains
  - Indentify impact of specific algorithm development across discipline groups
  - Take mapping of dwarfs or similar on capability computing
- Develop map of developments internationally
  i. Collect information about ongoing related activities
  ii. Discuss with international funding agencies what plans are in place in this area

A co-design approach is required

Theme 3: Software Challenges

There is a need for

- Abstractions (in collaboration with Computer Science) to allow more effective application development
- Code generation and adaptive software systems to automatically deliver efficient code for complex architectures
- Guidance on best practice for software engineering development
- Frameworks and tools for application developers to allow better reuse of algorithms
- Better understanding of usability issues for complex software systems

Theme 4: Sustainability

There is a need to:

- Address the sustainability of application codes, software libraries and skills
- develop models for sustainable HPC software that might include:
  - Long term funding
  - Industrial translation
  - Open community support

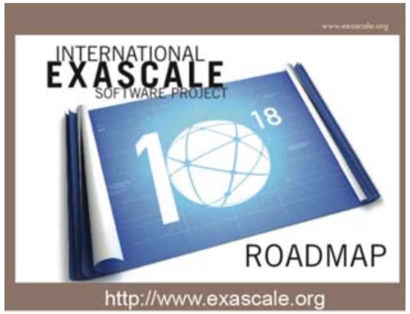
Theme 5: Knowledge Base

This theme is concerned with the general issue of sharing of knowledge and knowledge creation. The recommended actions are:

- Develop mechanisms for collecting information on existing software and expertise and dissemination
- Develop mechanism for continuing community input
- Develop appropriate education and training, through MScs, DTCs, short courses and summer schools.
- Engage industry, possibly through internships, to ensure industry needs are also met.

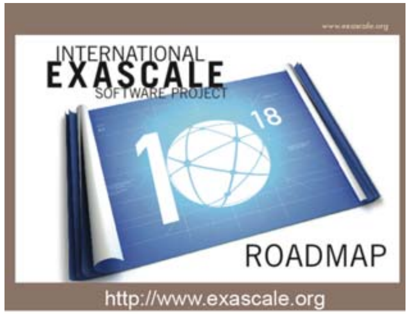
OXFORD
e-Research
CENTRE

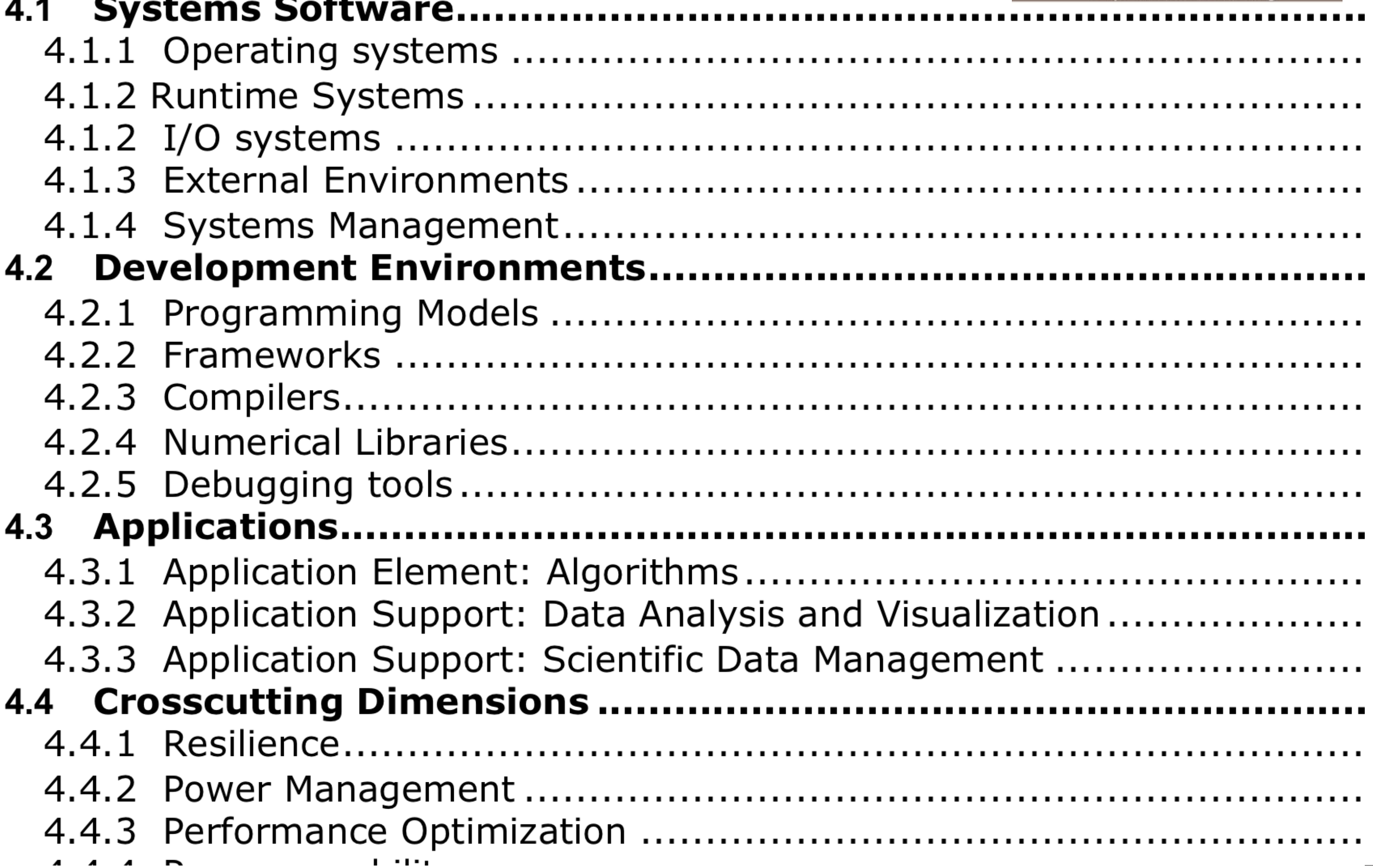
iesp findings



Key Trends	Requirements on X-Stack
❑ Increasing Concurrency	❑ **Programming models, applications and tools to address concurrency**
❑ Power Dominating designs	❑ **Power management by software**
❑ Reliability challenging	❑ **Resilience in software**
❑ Heterogeneity in a node	❑ **Software adapts to heterogeneity**
❑ I/O and Memory: ratios and breakthroughs	❑ **Software must be optimized for new memory ratios**

Roadmap in Draft





Co-Design Vehicles

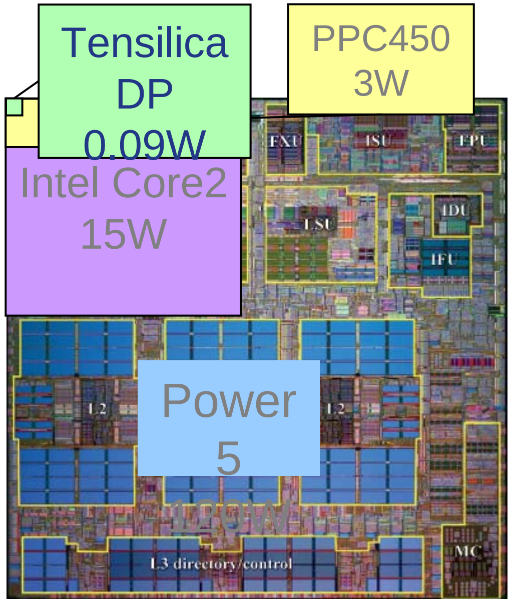
- ☐ Application/algorithm software & hardware development designed together to meet application needs

D.E Shaw

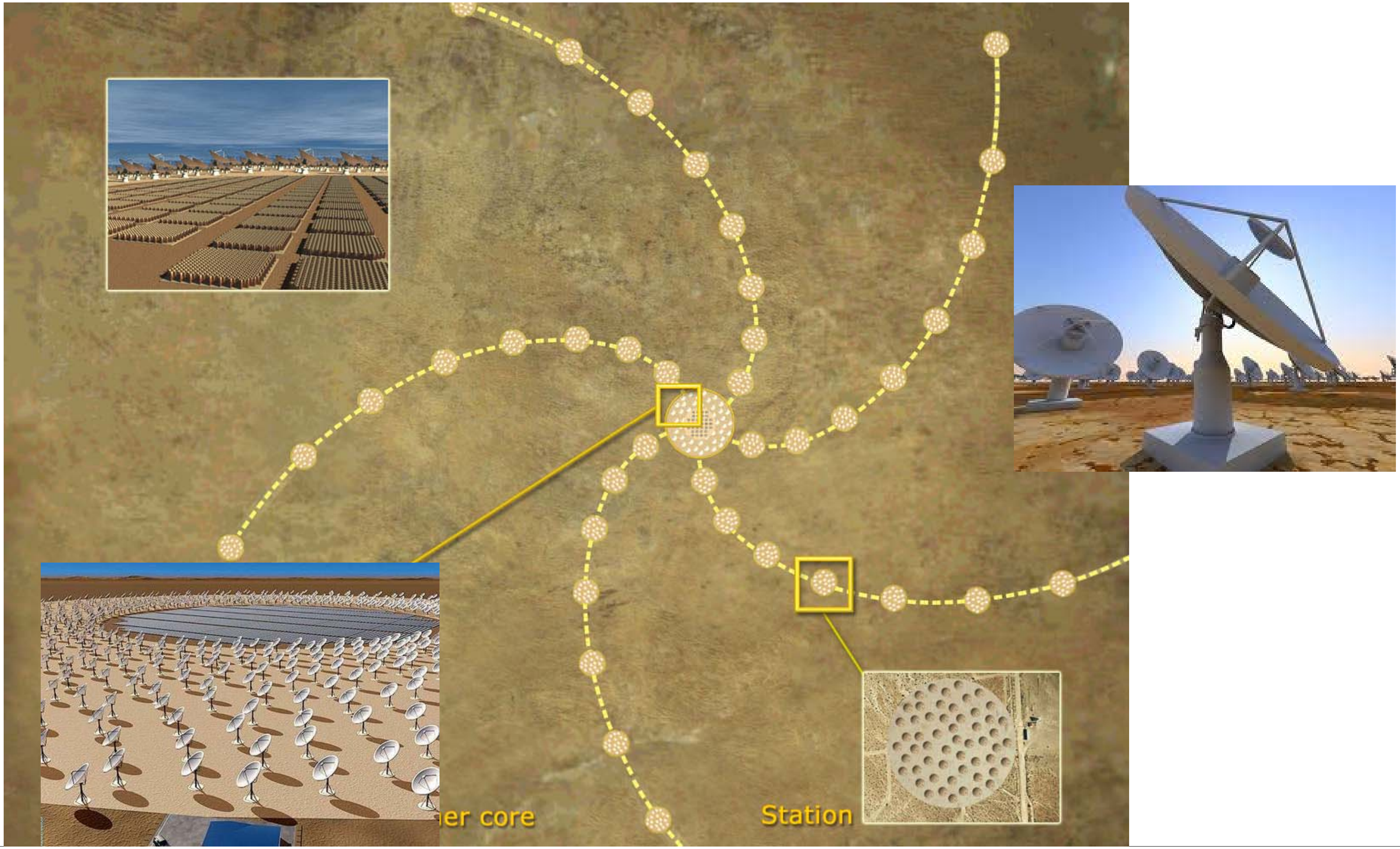
Green Flash

MD-Grape

QCDOC



Square Kilometre Array
Next Generation Radio Telescope



SKA: An Iconic Project

TECHNOLOGY

ICT (802.11A to Exascale computing)

“Will generate new ways of doing ICT that could revolutionize the world”
Bruce Elmegreen, IBM

GREEN COMPUTING (24/7 RE)

“Will play a global leadership role by aspiring to run 24/7 on Renewable Energy”
Eike Weber, Director Fraunhofer Institute

SCIENCE

“Will reveal profound truths about our Universe”
Steve Rawlings, Global Coordinator

Courtesy of Steve Rawlings

SKA Timeline

- 2012: PrepSKA delivers design for SKA$_1$, and Board makes site decision
- 2016-2019: SKA$_1$ construction & operation, and SKA$_2$ technology decision
- 2019-2022: SKA$_2$ construction & operation

OXFORD
e-Research
CENTRE

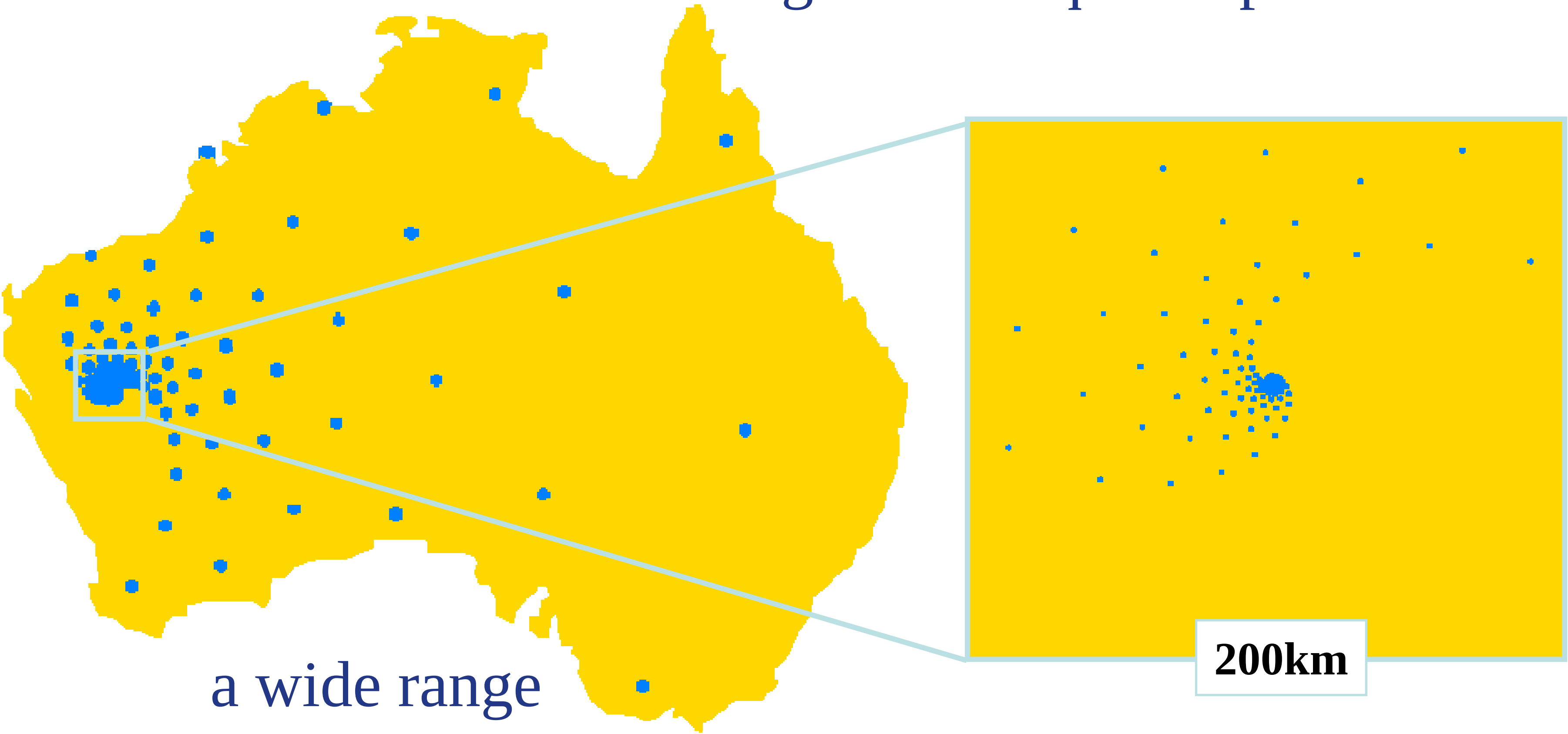
Science with the SKA

- ☐ The Universe in the Dark Ages
  - o Star formation
  - o epoch of (re-)ionization
- ☐ Cosmology and Large Scale Structure
  - o Gravitational Lensing
- ☐ Gamma Ray bursters
- ☐ AGN - VLBI
- ☐ Stellar radio astronomy
- ☐ Pulsars
- ☐ Solar system
- ☐ SETI

Individual science goals place different requirements on technology and algorithms

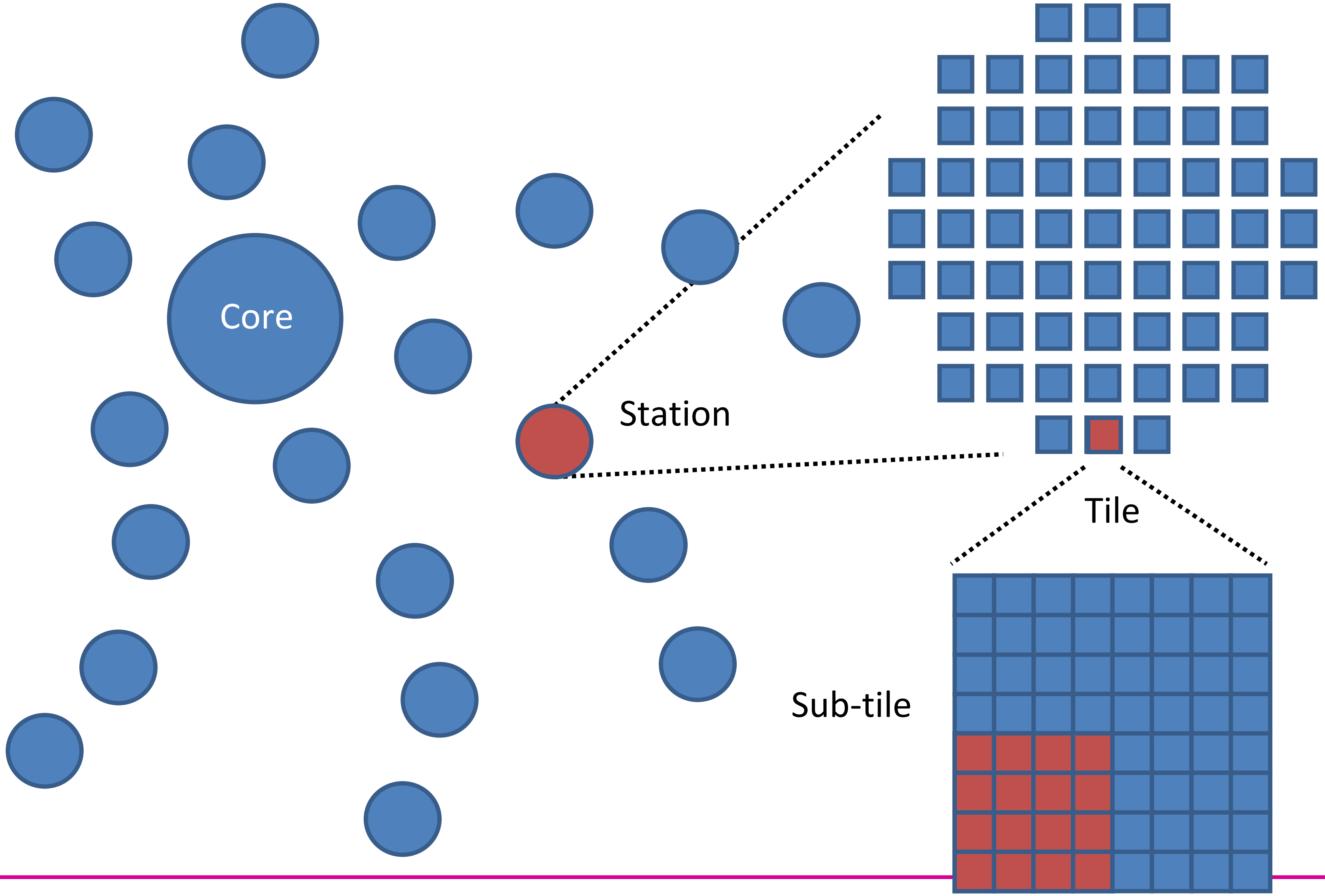
An example of a SKA configuration

Not a single 1 km square aperture !



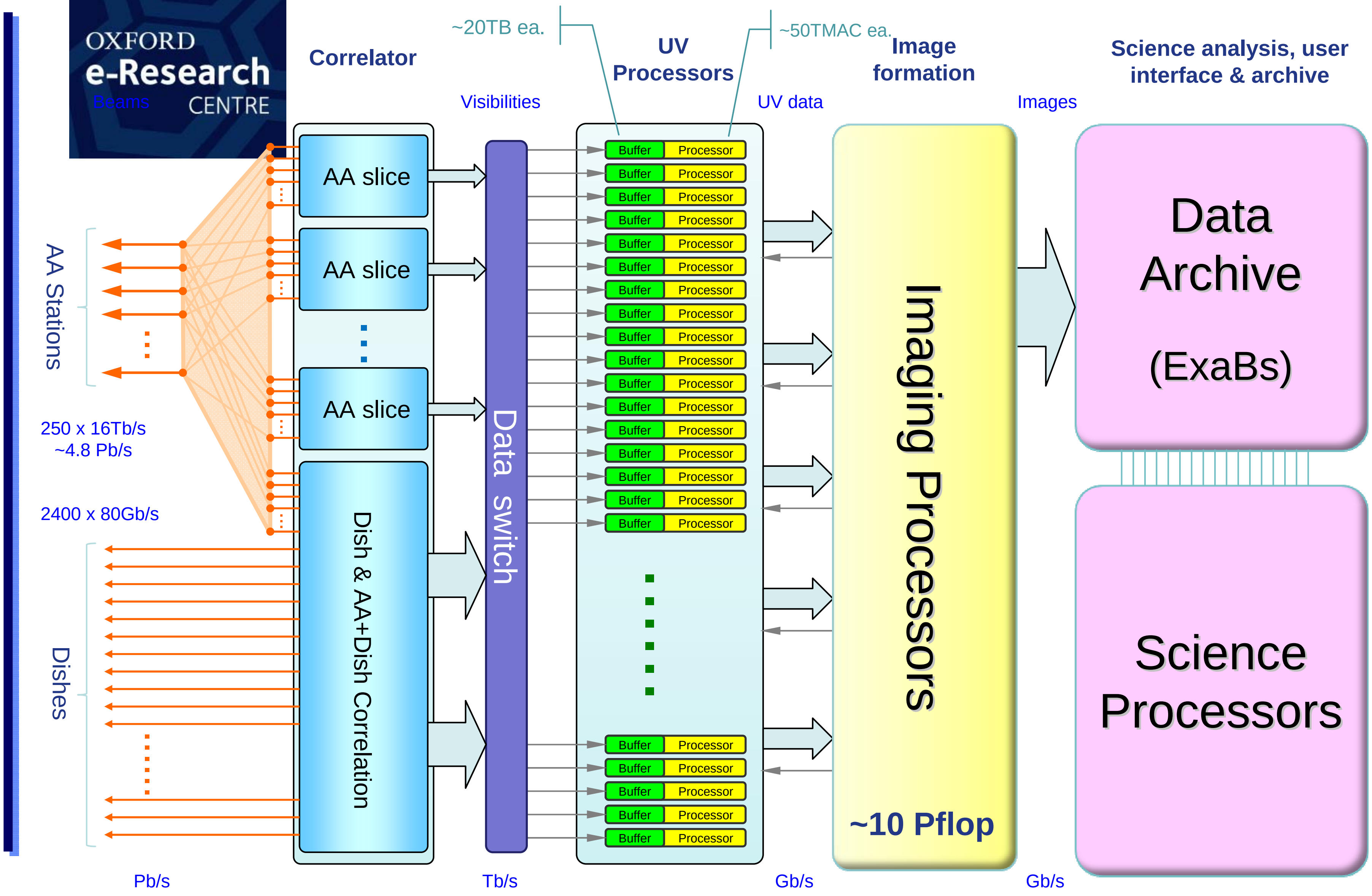
a wide range of baselines

SKA Structure



Central Processing Facility

OXFORD
e-Research
Beams CENTRE



~20TB ea.
~50TMAC ea.

Correlator
UV Processors
Image formation
Science analysis, user interface & archive

Visibilities
UV data
Images

AA Stations
250 x 16Tb/s
~4.8 Pb/s
2400 x 80Gb/s
Dishes

AA slice
AA slice
AA slice
Dish & AA+Dish Correlation

Data switch

Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor
Buffer Processor

Imaging Processors
~10 Pflop

Data Archive
(ExaBs)

Science Processors

Pb/s
Tb/s
Gb/s
Gb/s

Courtesy of Steve Rawlings

ICT Challenges for SKA

Electricity costs ~€0.2 per kW hr
or ~€70M each year

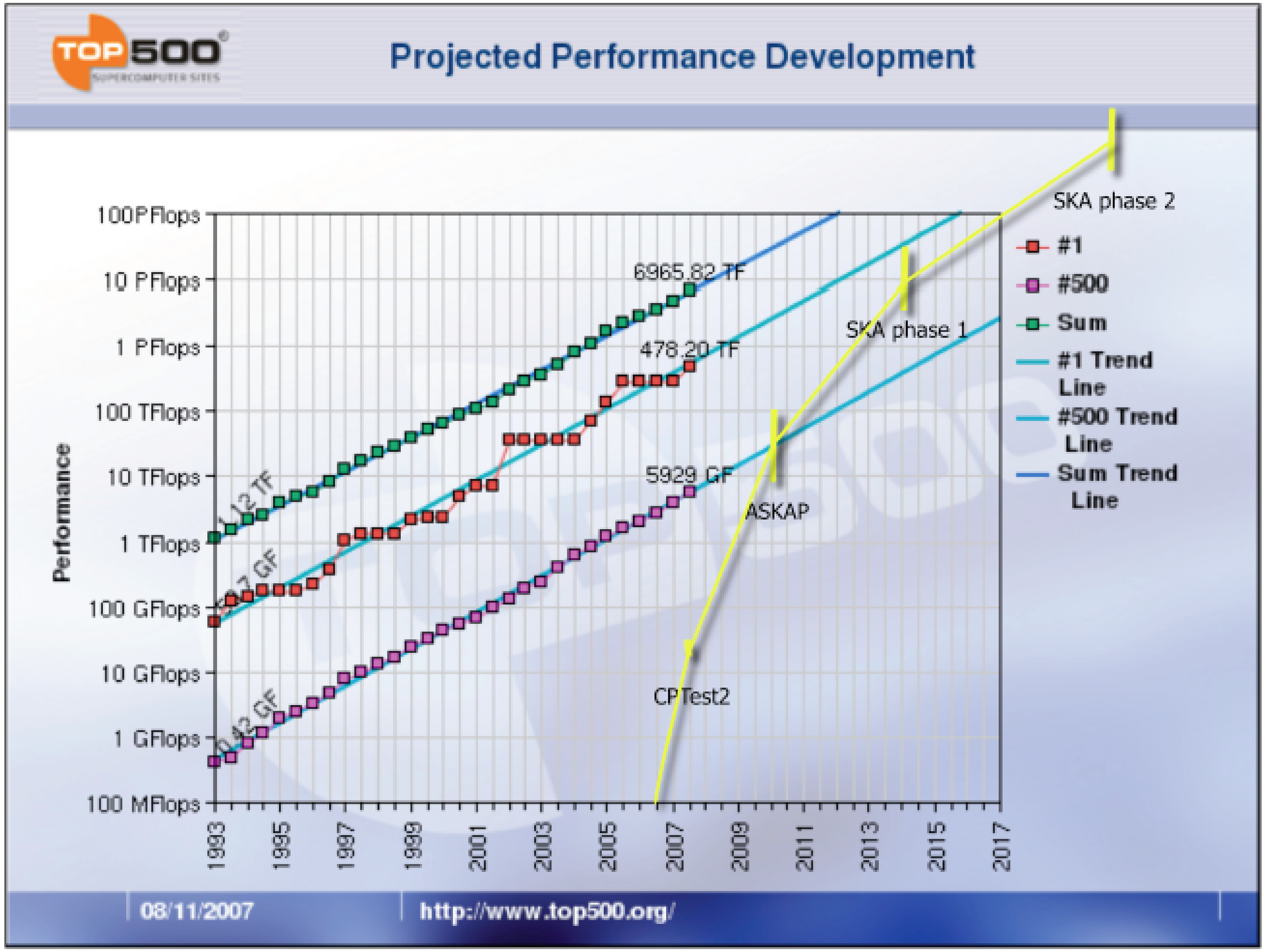
1 kW m^{-2} onto ~1 km^2 @ 10% efficiency can delivers ~100 MW: power requirement of SKA

- > 10 Tb/s network + “Mount ExaFlop”
- ~30000 40 TMACs DSP engines
- ~10000 50-Tflop many-core processors
- >10 Pflop supercomputer
- Pb/s input to ExaByte archive
- ~100 MW power budget

Algorithms include FFT, Correlation, Filters, etc

Courtesy of Steve Rawlings

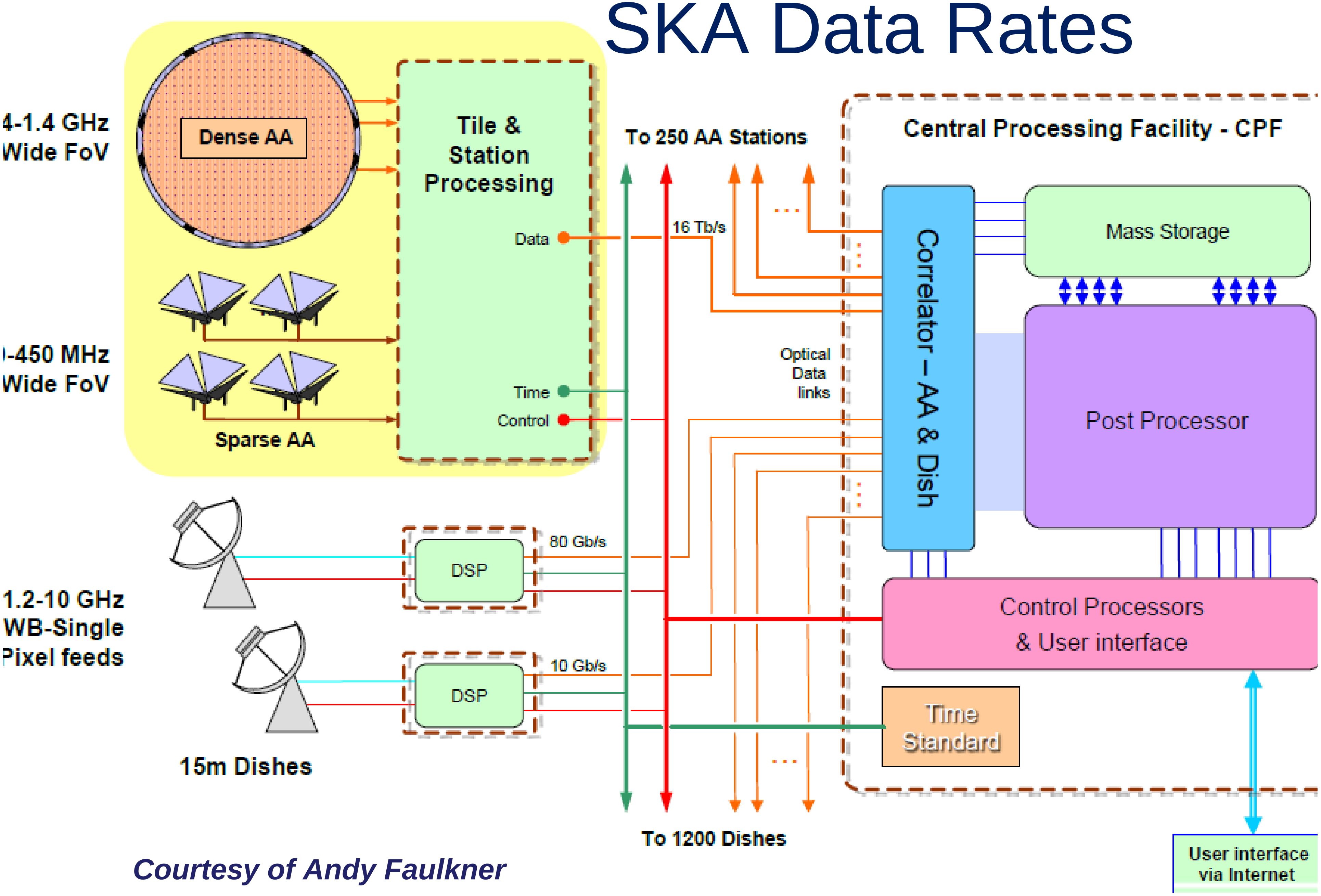
Scaling Mount Exaflop



Courtesy of
Tim Cornwell

ure 3 Expected growth of processing requirements for ASKAP and SKA. CPTest2 is our first test of

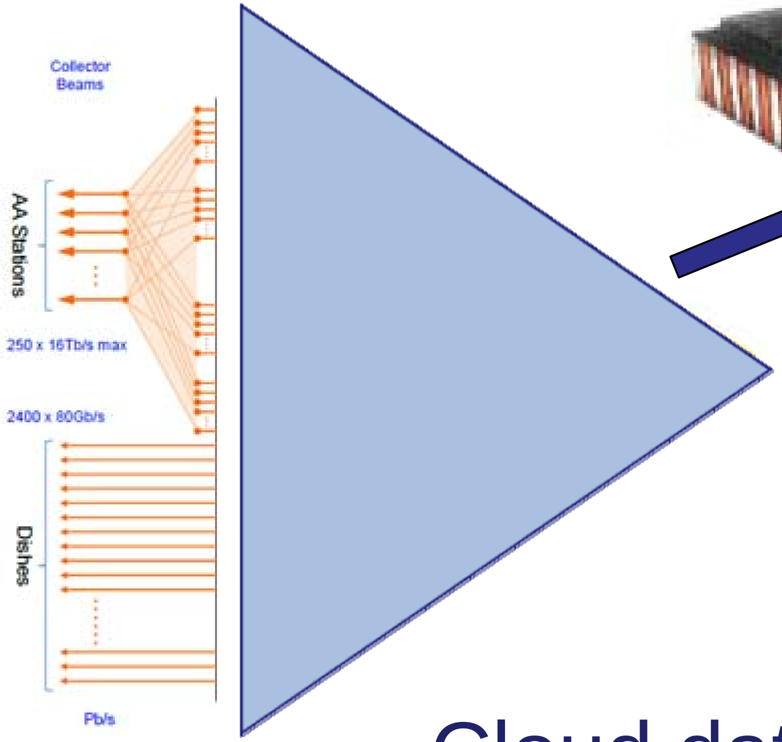
SKA Data Rates



Courtesy of Andy Faulkner

OXFORD
e-Research
CENTRE

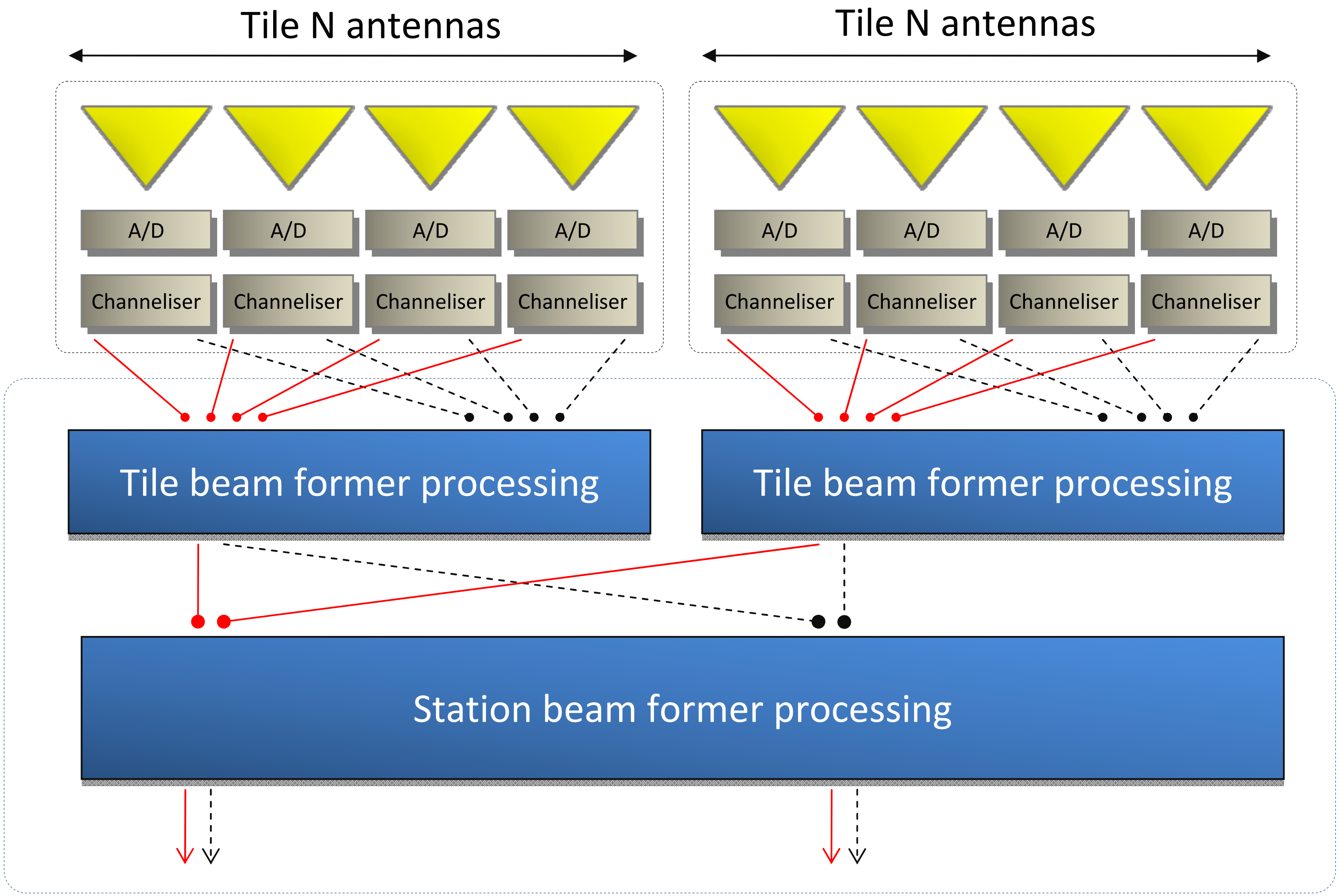
The computing ecosystem for SKA



Exaflop

Cloud data services

Simple Beamforming Scheme



Towards a Strategy

- **Hierarchical beam forming**
    - Tile: 16 x 16 (256 element, dual polarisation) matrix of antennas.
    - Sub-tile?: 8 x 8 quadrant of tile.
    - Sub-stations?
- **Tile beams combined as required to give station beam.**
- **Questions:**
    - How many antennas per tile?
    - How many tiles per station?
    - Do we need sub-tiles
    - Overlapping tiles?

Flops & Gigaflops

- **Assumptions**
  - “Full matrix” filter, applicable to beams space
  - Cheaper if filters expressible as Cartesian product of filters along each dimension
  - Disjoint tiles (non overlapping)
- **Main conclusions**
  - Cost linearly dependent on n. tile beams x n. channels

	Filter	Value
n	Tile size	16
m	Station size in tiles	16
b	N. Beams from tile level	1
B	N. Beams at station level	256
	Sampling rate (after channeliser)	1 Mhz
	Number of channels	1024

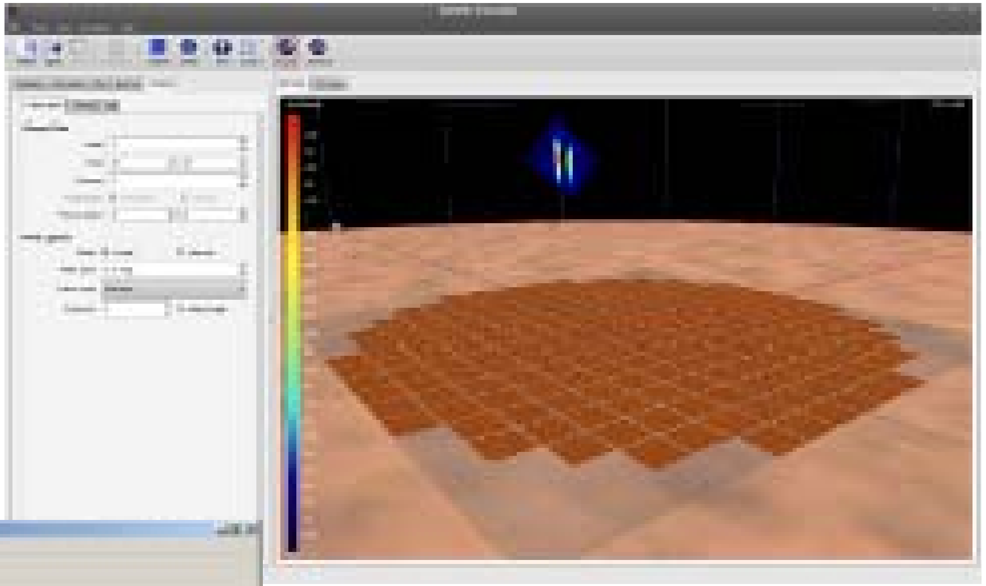
Scheme	Filter	N. Flops per channel
FFT	Full Matrix (beam)	$b^2 \cdot (2 \cdot 4 \cdot n^2 \cdot \log_2 n + 8 \cdot n^2)$
Full-Matrix	Any	$8 \cdot b^2 \cdot n^2$

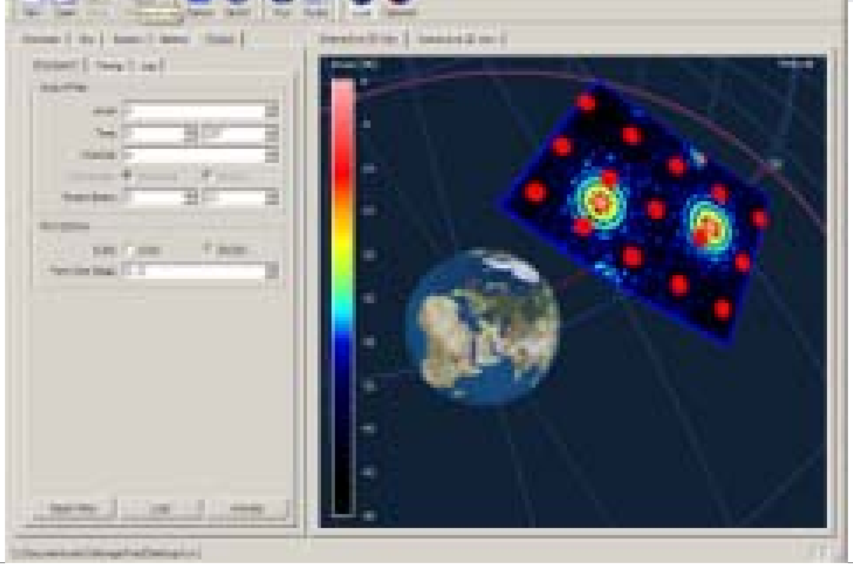
Computational Costs per Station (1 tile beam)			
Scheme	**Filter**	**Tiles Gflops/channel**	**Station GFlops per channel**
FFT	Full Matrix	2130	2660
Full-Matrix	Any	525	1050

Efficiency = Simple Code

- **Limited set of operations**
  - Matrix-vector & matrix-matrix products only?
- **Weights matrices (DFT * Filters)**
  - Vary much slower than the sampling rate (10,000 slower?)
  - May be computed offline on modest computing resources
  - Tables, using then high-order interpolant?
- **“Easy” to port to different hardware architectures**
  - Linear striding through memory
  - Possible to design bespoke chips while retaining maximum flexibility
- **“Trivial” parallelism**
  - No interprocessor communication during computation

DSP Beam-forming: OSKAR



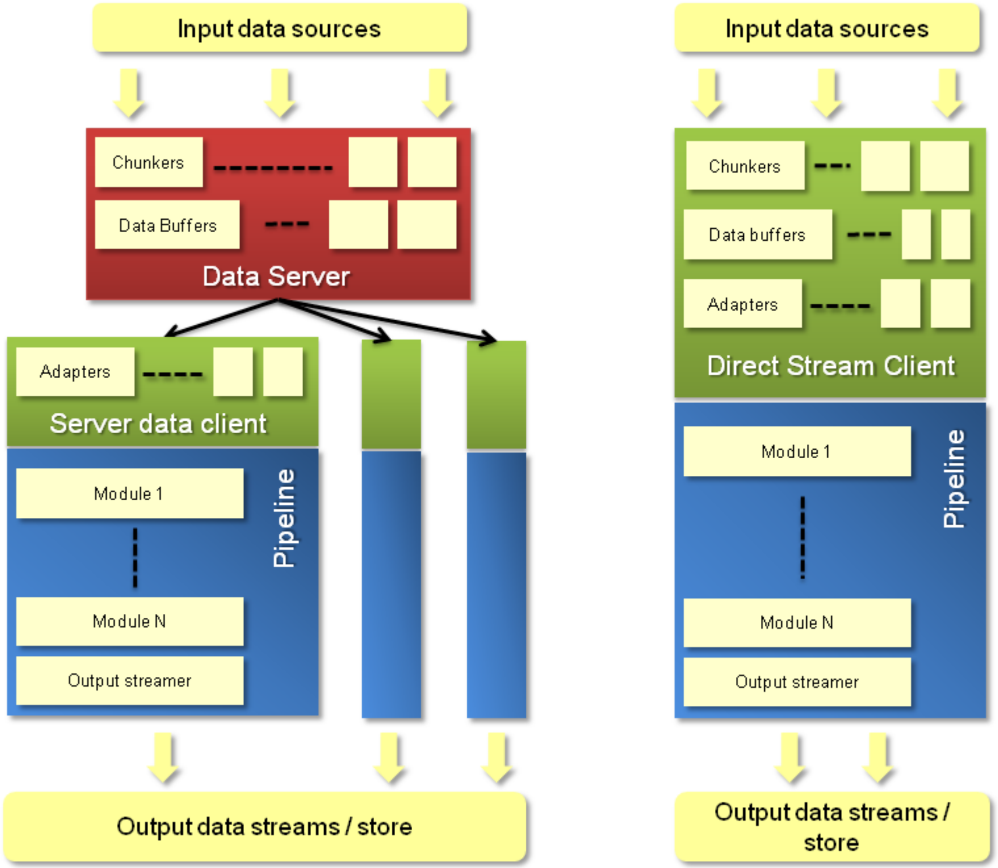


Can now simulate ~1s of SKA station data!

Software open source and available on oskar wiki

The Pelican Framework

- Pelican is a C++ framework for parallel quasi-real time data processing.
- Two deployment options.
  - Server supplies multiple pipelines.
  - Pipeline connects directly to data stream.
- Server and pipelines are constructed from reusable modular components.



Input data sources

Chunkers

Data Buffers

Data Server

Adapters

Server data client

Module 1

Module N

Output streamer

Pipeline

Output data streams / store

Input data sources

Chunkers

Data buffers

Adapters

Direct Stream Client

Module 1

Module N

Output streamer

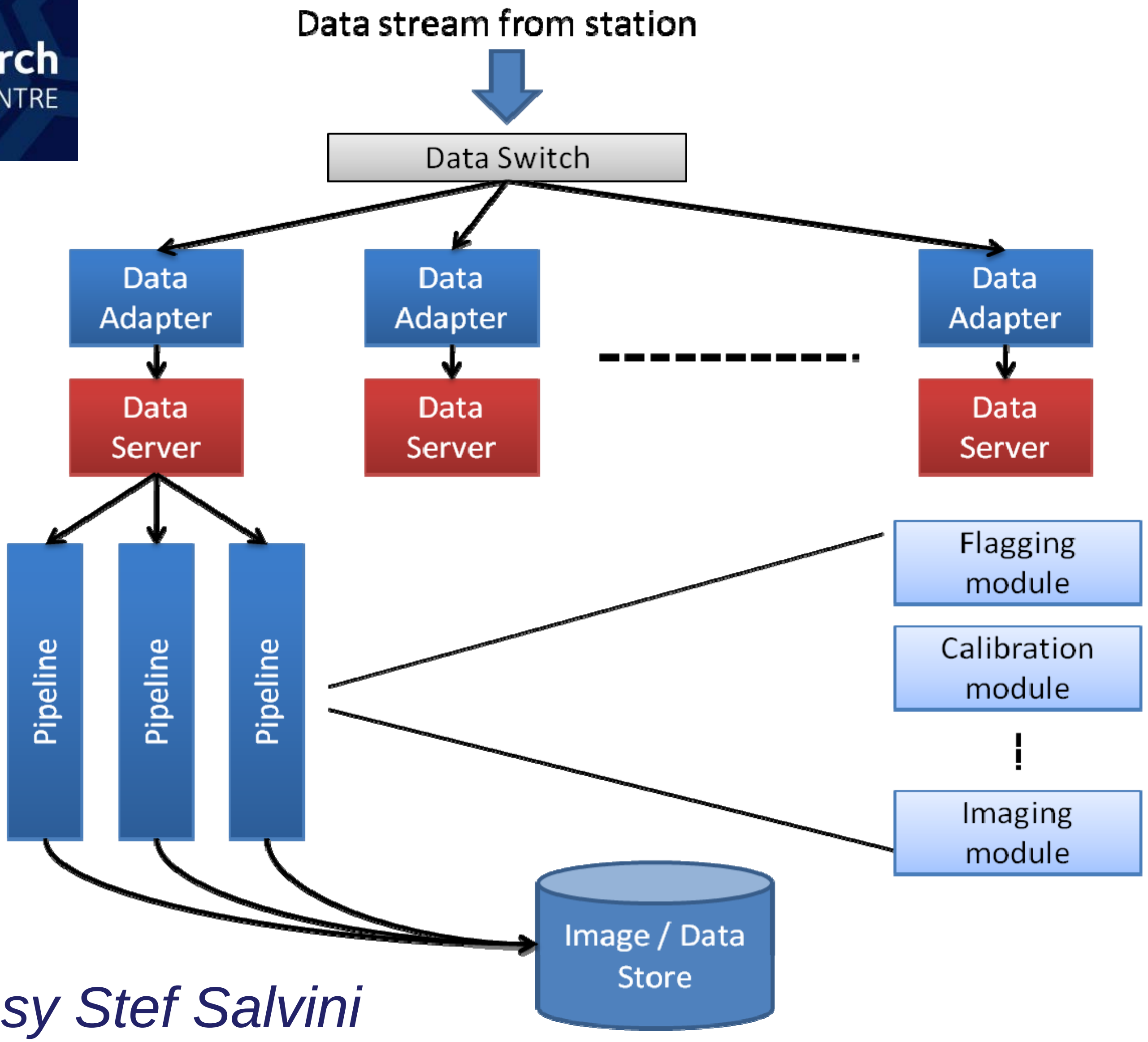
Pipeline

Output data streams / store

Courtesy Stef Salvini

OXFORD
e-Research
CENTRE

STREAM PROCESSING: PELICAN



Courtesy Stef Salvini

Using Pelican

- Used for processing radio astronomical data in real time.
- To be deployed on LOFAR interferometer stations:
    - All sky calibration and imaging.
    - Pre-processing for pulsar searching.
- Input data rate of 3.2 Gb/s

Press Release

OXFORD
e-Research
CENTRE

An Oxford-based team has played a key role in creating software with the potential to change the way astronomers look at the Universe.
By linking up a next-generation radio observatory, general purpose computer graphics chips and some highly sophisticated software for handling large volumes of streaming data, their work will help astronomers observe and study some of the most extreme events ever known......

The first actually operational implementation!

SKA ICT Design features

- Need ASIC or FPGA-like device close to antenna
- Exaflop computation
- Provision of large-scale data archive and services
- Constrained by power, costs, technology capability

Energy efficiency

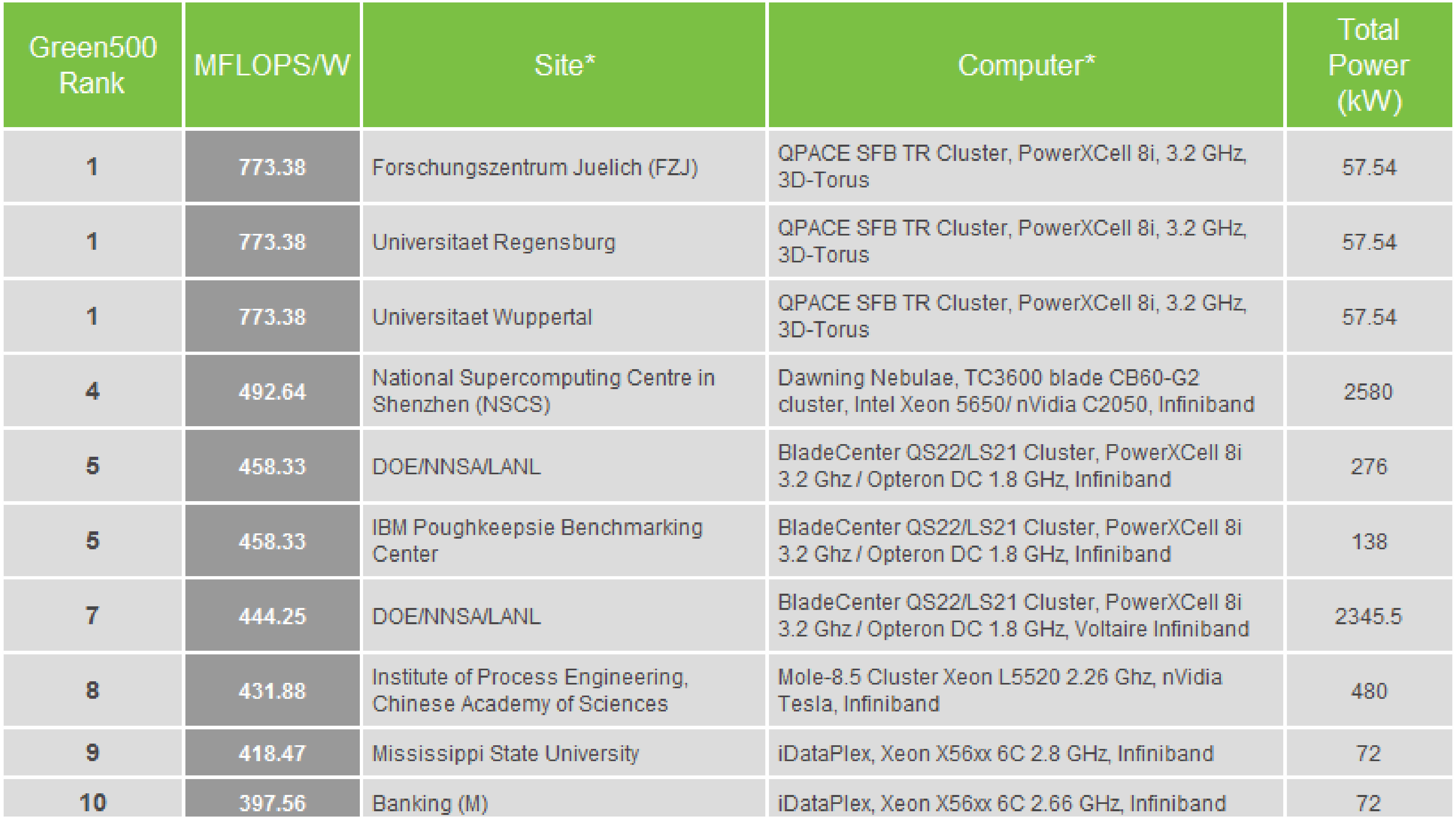
- Need to have energy efficiency at every step
    - Antenna
    - Data communication
    - Exascale computation
    - Cloud computation
    - Desktop analysis

- Power trends in computing
    - We have seen about a 2.5x system level power efficiency improvement over the last 3 years.
    - We need about 100x improvement over the next 10 years to get to a 20 MW Exaflop system.

OXFORD e-Research CENTRE

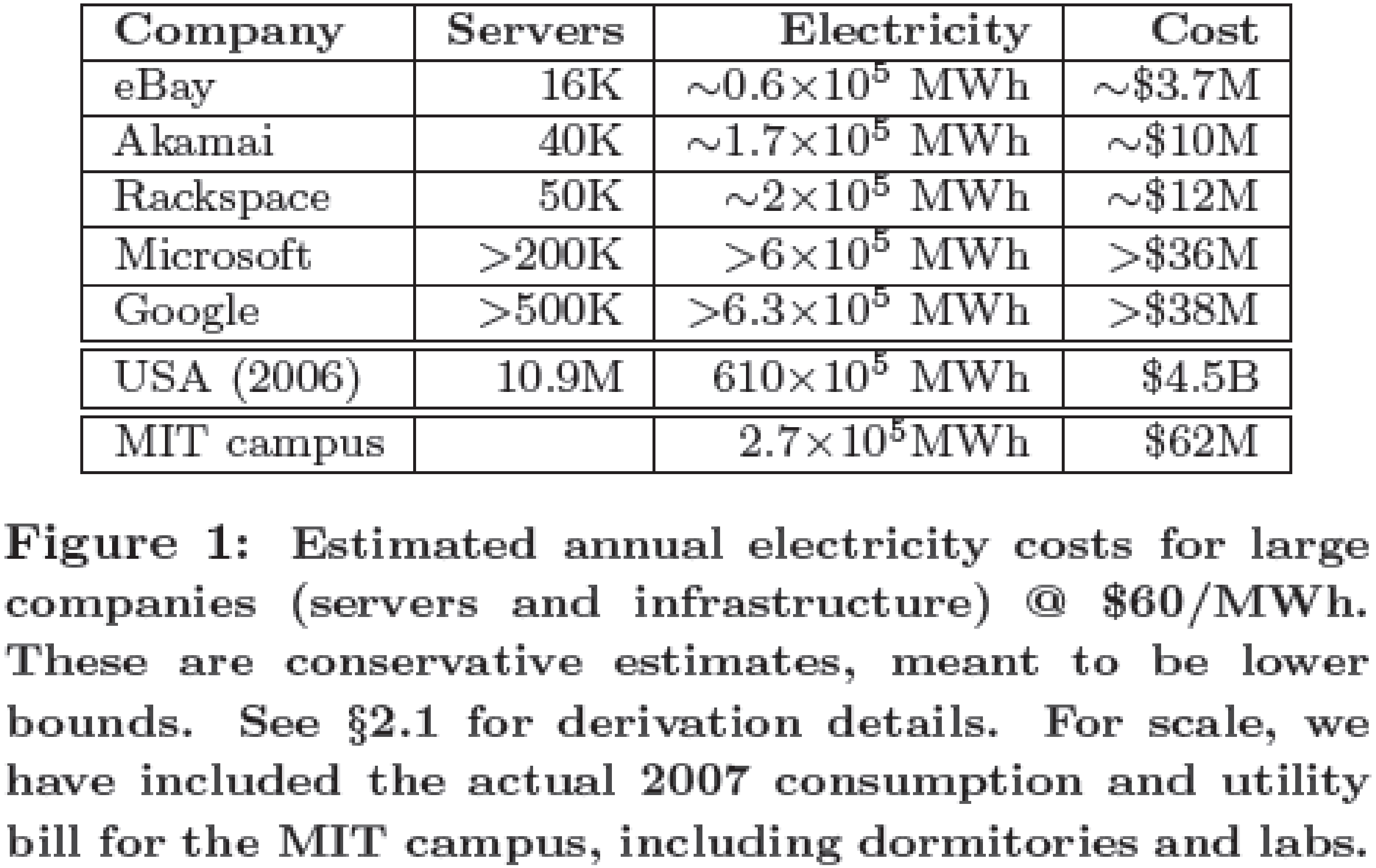
www.green500.org

THE GREEN 500



Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	773.38	Forschungszentrum Juelich (FZJ)	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
1	773.38	Universitaet Regensburg	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
1	773.38	Universitaet Wuppertal	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
4	492.64	National Supercomputing Centre in Shenzhen (NSCS)	Dawning Nebulae, TC3600 blade CB60-G2 cluster, Intel Xeon 5650/ nVidia C2050, Infiniband	2580
5	458.33	DOE/NNSA/LANL	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Infiniband	276
5	458.33	IBM Poughkeepsie Benchmarking Center	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Infiniband	138
7	444.25	DOE/NNSA/LANL	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband	2345.5
8	431.88	Institute of Process Engineering, Chinese Academy of Sciences	Mole-8.5 Cluster Xeon L5520 2.26 Ghz, nVidia Tesla, Infiniband	480
9	418.47	Mississippi State University	iDataPlex, Xeon X56xx 6C 2.8 GHz, Infiniband	72
10	397.56	Banking (M)	iDataPlex, Xeon X56xx 6C 2.66 GHz, Infiniband	72

Cloud computing energy costs?



Company	Servers	Electricity	Cost
eBay	16K	$\sim 0.6 \times 10^5$ MWh	~$3.7M
Akamai	40K	$\sim 1.7 \times 10^5$ MWh	~$10M
Rackspace	50K	$\sim 2 \times 10^5$ MWh	~$12M
Microsoft	>200K	$> 6 \times 10^5$ MWh	>$36M
Google	>500K	$> 6.3 \times 10^5$ MWh	>$38M
USA (2006)	10.9M	610×10^5 MWh	$4.5B
MIT campus		2.7×10^5MWh	$62M

Figure 1: Estimated annual electricity costs for large companies (servers and infrastructure) @ $60/MWh. These are conservative estimates, meant to be lower bounds. See §2.1 for derivation details. For scale, we have included the actual 2007 consumption and utility bill for the MIT campus, including dormitories and labs.

From Cutting the Electric Bill for Internet-Scale Systems, Qureshi et al.

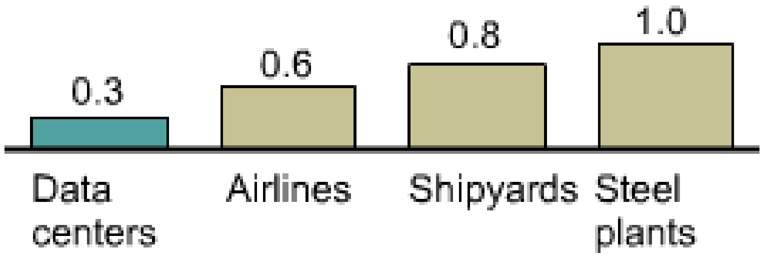
OXFORD e-Research CENTRE

McKinsey/Uptime report on Data Centres

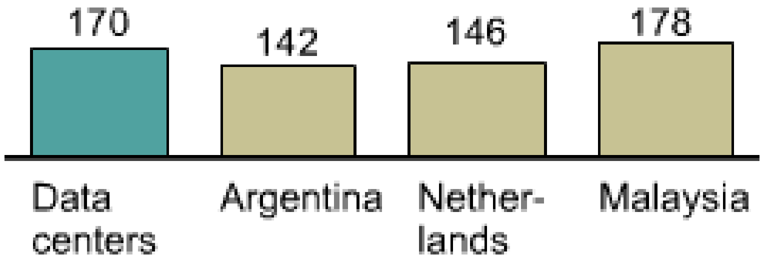
Key points on data centers' greenhouse gas emissions

- Data center **electricity consumption** is **almost .5% of world production***
- Average data center consumes energy equivalent to 25,000 households
- Worldwide energy consumption of DC doubled between 2000 and 2006
- Incremental US demand for data center energy between now and 2010 is equivalent of 10 new power plants
- 90% of companies running large data centers need to build more power and cooling in the next 30 months

Carbon dioxide emissions as percentage of world total – industries
Percent



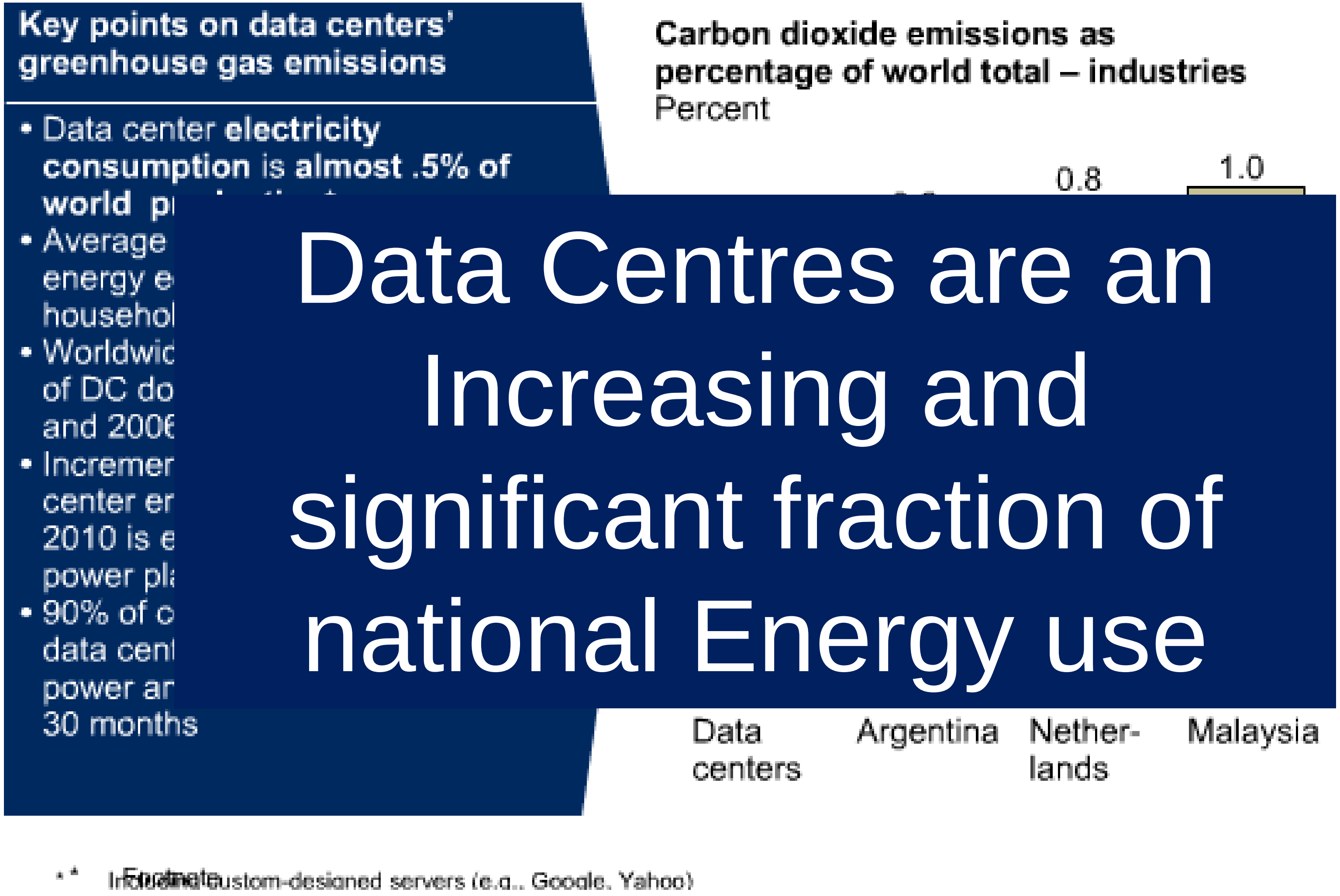
Carbon emissions – countries
Mt CO_2 p.a.



Footnote

* Includes custom-designed servers (e.g., Google, Yahoo)

McKinsey/Uptime report on Data Centres



Key points on data centers' greenhouse gas emissions

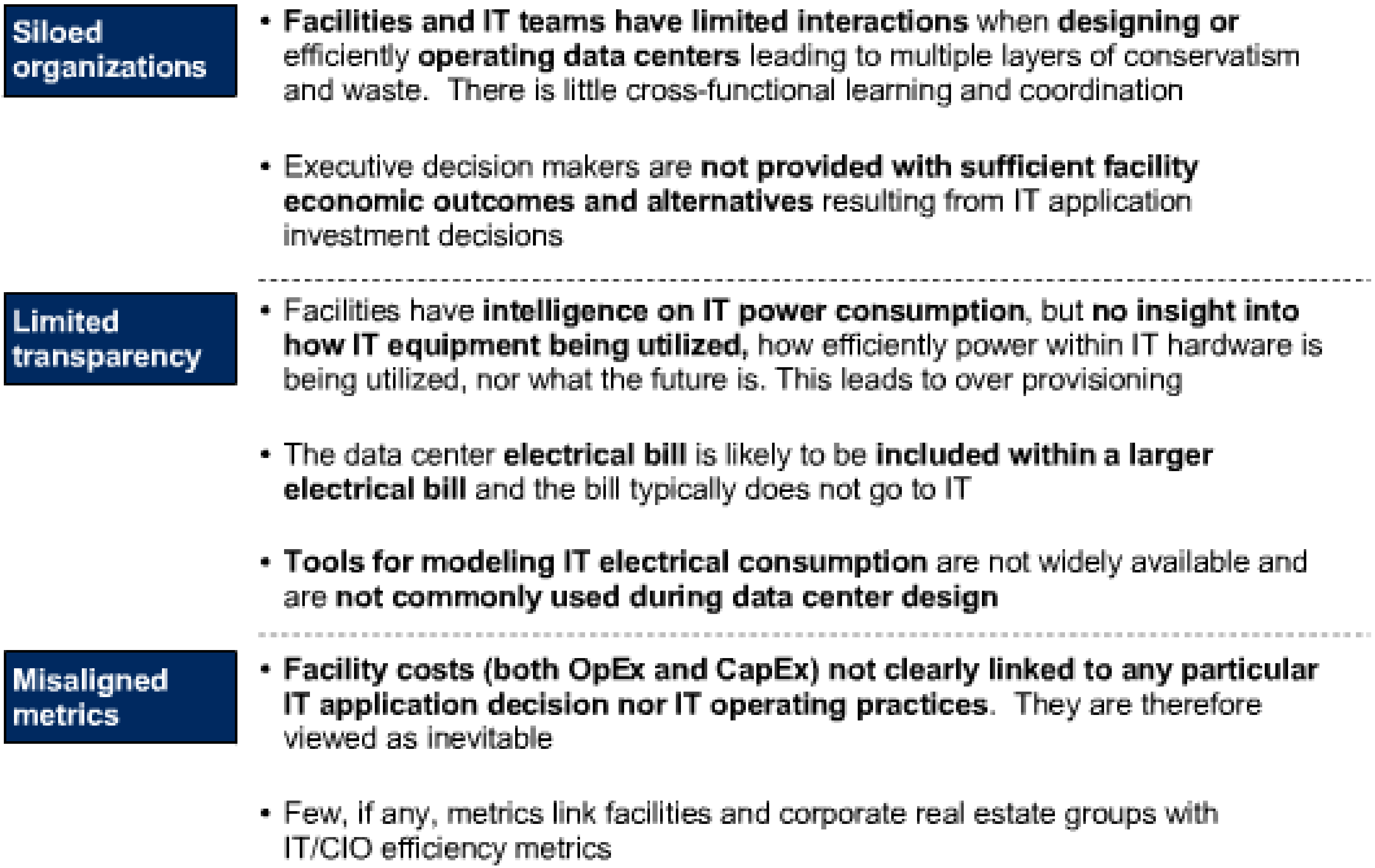
- Data center **electricity consumption** is **almost .5% of world** p
- Average energy e househol
- Worldwid of DC do and 2006
- Incremen center en 2010 is e power pl
- 90% of c data cent power ar 30 months

Carbon dioxide emissions as percentage of world total – industries
Percent

* * Footnote Includes custom-designed servers (e.g., Google, Yahoo)

Data Centres are an Increasing and significant fraction of national Energy use

McKinsey Findings



Siloed organizations

- **Facilities and IT teams have limited interactions** when **designing or** efficiently **operating data centers** leading to multiple layers of conservatism and waste. There is little cross-functional learning and coordination
- Executive decision makers are **not provided with sufficient facility economic outcomes and alternatives** resulting from IT application investment decisions

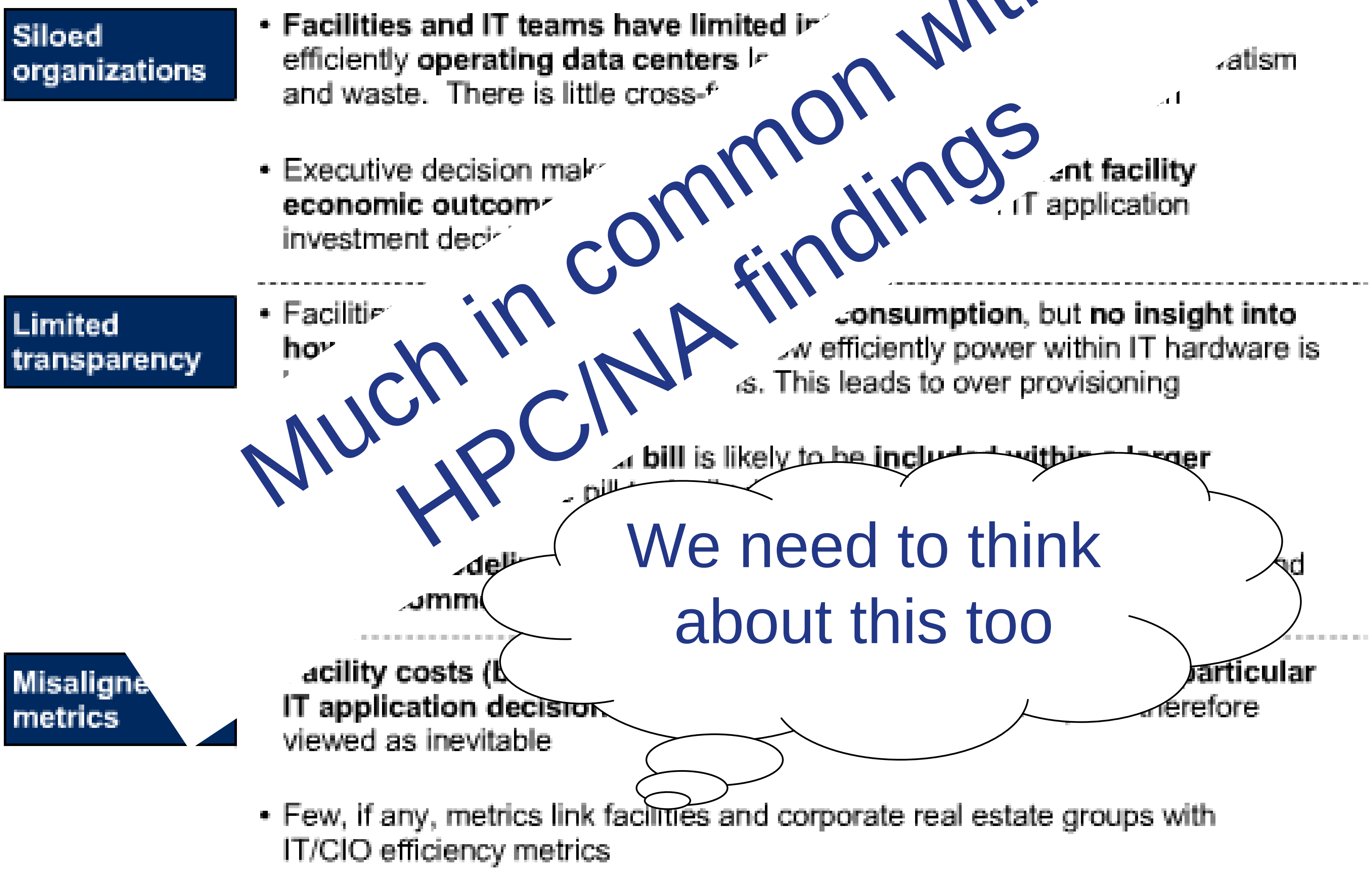
Limited transparency

- Facilities have **intelligence on IT power consumption**, but **no insight into how IT equipment being utilized,** how efficiently power within IT hardware is being utilized, nor what the future is. This leads to over provisioning
- The data center **electrical bill** is likely to be **included within a larger electrical bill** and the bill typically does not go to IT
- **Tools for modeling IT electrical consumption** are not widely available and are **not commonly used during data center design**

Misaligned metrics

- **Facility costs (both OpEx and CapEx) not clearly linked to any particular IT application decision nor IT operating practices**. They are therefore viewed as inevitable
- Few, if any, metrics link facilities and corporate real estate groups with IT/CIO efficiency metrics

McKinsey Find



Siloed organizations

- **Facilities and IT teams have limited in**
efficiently **operating data centers** l ... atism
and waste. There is little cross-f
- Executive decision mak ... **ent facility**
economic outcom ... IT application
investment deci

Limited transparency

- Faciliti ... **consumption**, but **no insight into**
how ... w efficiently power within IT hardware is
... s. This leads to over provisioning
... **bill** is likely to be **included within a larger**
... bill
odeli ... nd
omm

Misaligned metrics

- **acility costs (** ... **particular**
IT application decision ... therefore
viewed as inevitable
- Few, if any, metrics link facilities and corporate real estate groups with IT/CIO efficiency metrics

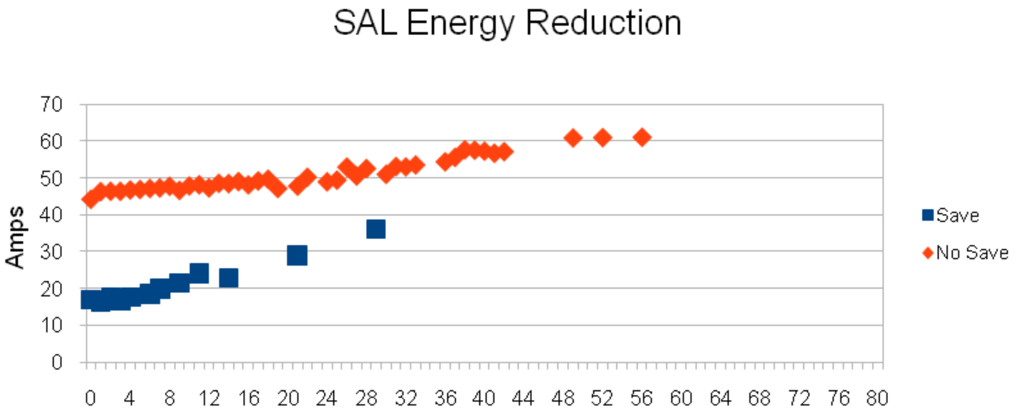
Much in common with HPC/NA findings

We need to think about this too

OXFORD
e-Research
CENTRE

At Oxford Supercomputing Centre

- Software queries the job scheduler as to the state of the cluster
- If a node is empty and powered up, an “Action” is taken
- If a node is powered down (or in some other user defined state), another “Action” is taken
- Actions can be applied to nodes in sequence or at random
- Actions are applied at a user defined interval
- Actions are user defined
- All cluster states and actions are logged in a SQL database



Courtesy of Jon Lockley

OXFORD
e-Research
CENTRE

Green Grid: Creating standards

- The Green Grid is a global consortium dedicated to developing and promoting energy efficiency for data centers and business computing ecosystems by:
  - Defining meaningful, user-centric models and metrics
  - Promoting the adoption of energy efficient standards, processes, measurement methods and technologies
  - Developing standards, measurement methods, processes and new technologies to improve performance against the defined metrics

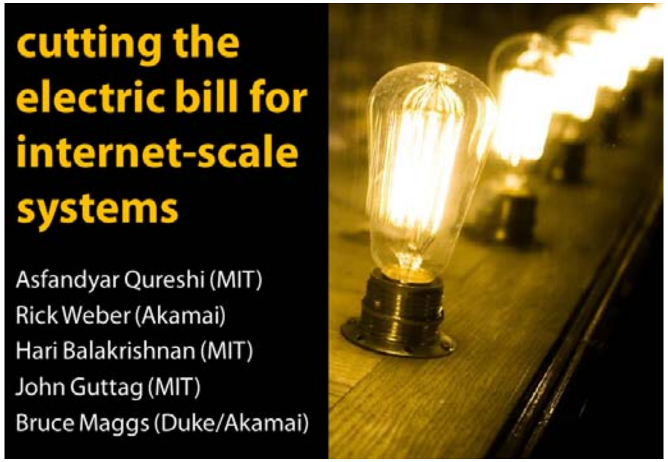
OXFORD
e-Research
CENTRE

Energy-aware communications

INTERNET
INTelligent Energy awaRe NETworks

Lot of research on energy – aware and efficient sensor and wireless networks

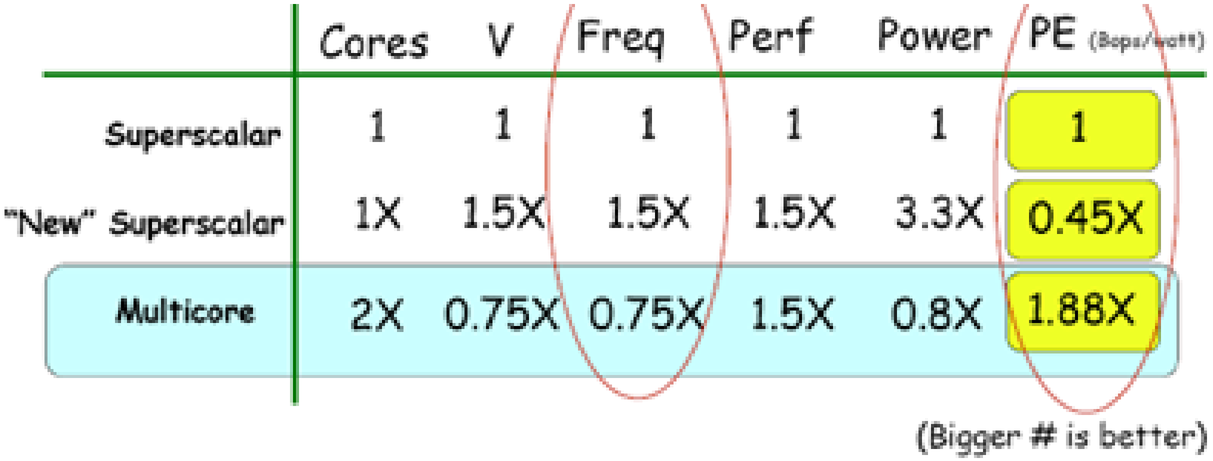
Number of projects now looking at optimising energy cost for Internet based activity


cutting the electric bill for internet-scale systems

Asfandyar Qureshi (MIT)
Rick Weber (Akamai)
Hari Balakrishnan (MIT)
John Guttag (MIT)
Bruce Maggs (Duke/Akamai)

Power Cost of devices

- o Power α Voltage^2 x Frequency (V^2F)
- o Frequency α Voltage
- o Power α Frequency ^{3}

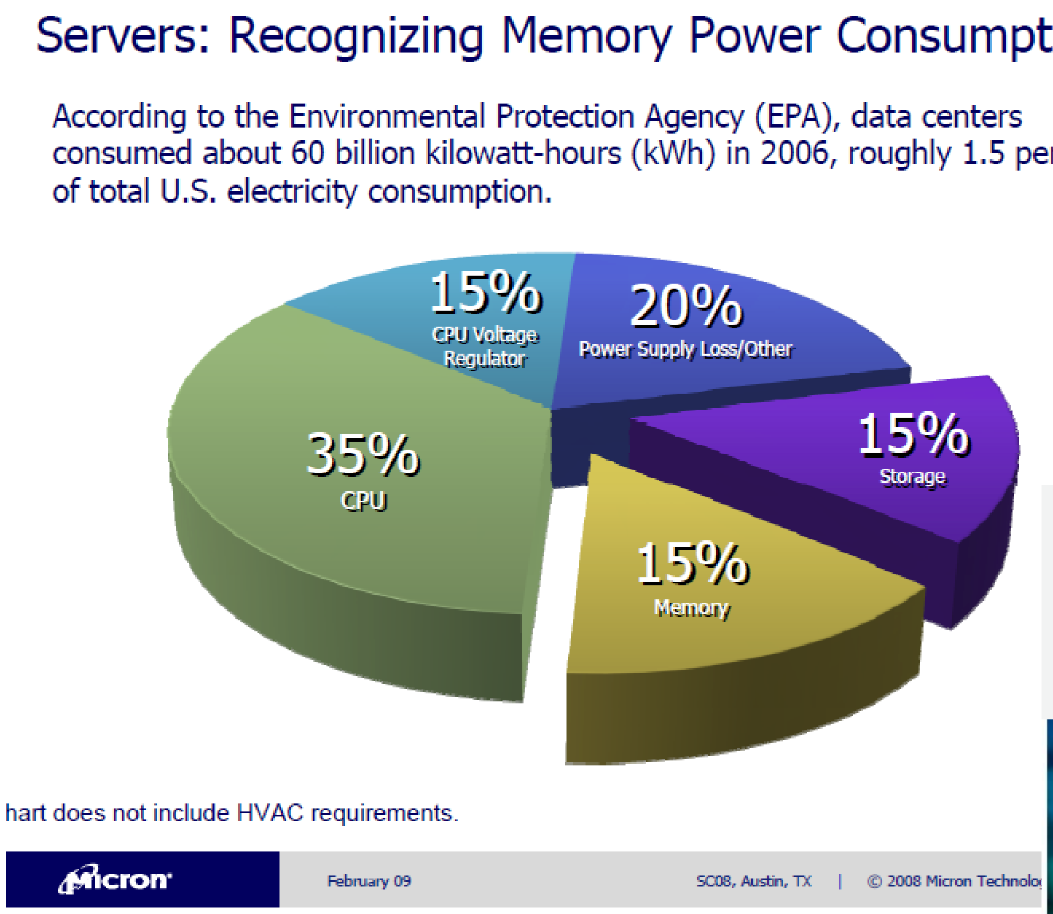


	Cores	V	Freq	Perf	Power	PE (Bops/watt)
Superscalar	1	1	1	1	1	1
"New" Superscalar	1X	1.5X	1.5X	1.5X	3.3X	0.45X
Multicore	2X	0.75X	0.75X	1.5X	0.8X	1.88X

(Bigger # is better)

Multicore 50% more performance with 20% less power

3D memory stacking on multiprocessors & on-chip communications

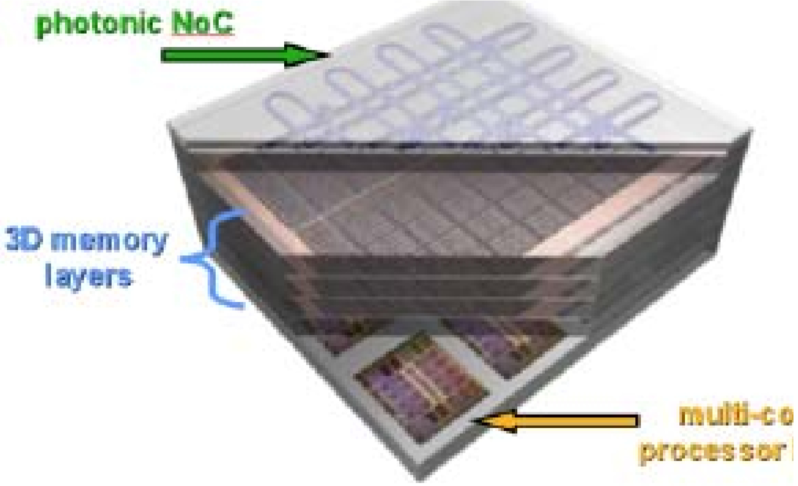


Servers: Recognizing Memory Power Consumption

According to the Environmental Protection Agency (EPA), data centers consumed about 60 billion kilowatt-hours (kWh) in 2006, roughly 1.5 percent of total U.S. electricity consumption.

hart does not include HVAC requirements.

Micron February 09 SC08, Austin, TX | © 2008 Micron Technolo

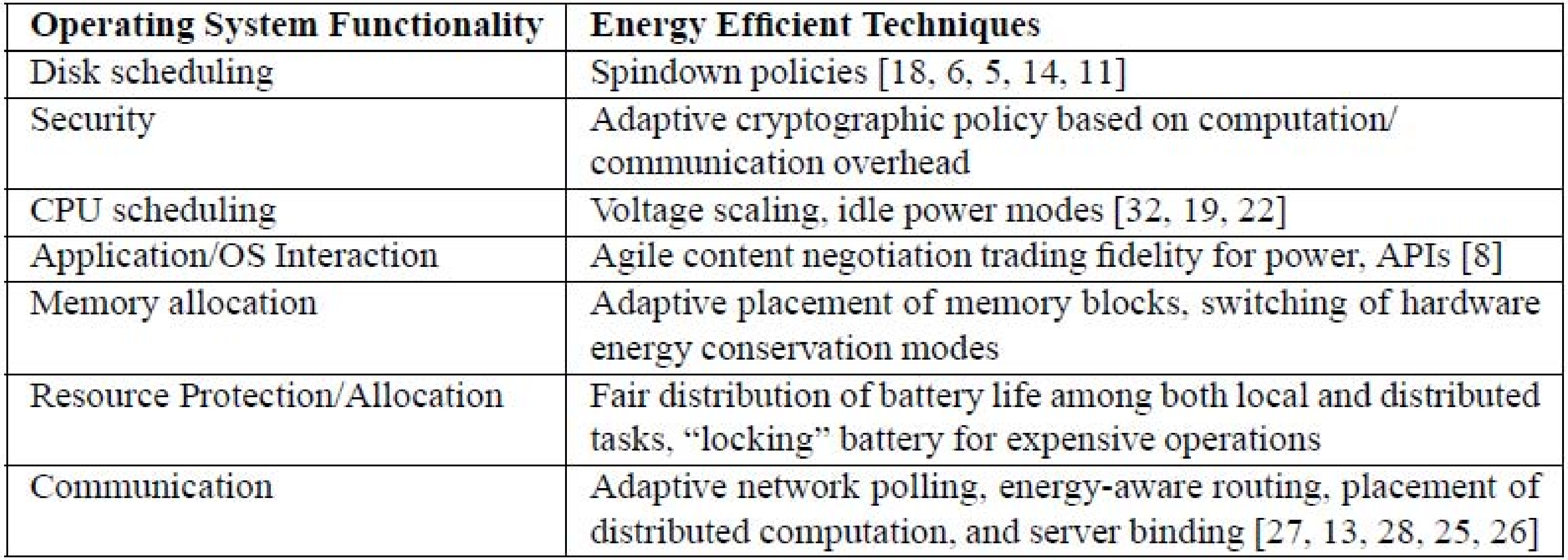


Intel demos 50-Gbit/s silicon optics

R. Colin Johnson

7/27/2010 1:30 PM EDT

Efficiencies at Operating system



Operating System Functionality	Energy Efficient Techniques
Disk scheduling	Spindown policies [18, 6, 5, 14, 11]
Security	Adaptive cryptographic policy based on computation/ communication overhead
CPU scheduling	Voltage scaling, idle power modes [32, 19, 22]
Application/OS Interaction	Agile content negotiation trading fidelity for power, APIs [8]
Memory allocation	Adaptive placement of memory blocks, switching of hardware energy conservation modes
Resource Protection/Allocation	Fair distribution of battery life among both local and distributed tasks, "locking" battery for expensive operations
Communication	Adaptive network polling, energy-aware routing, placement of distributed computation, and server binding [27, 13, 28, 25, 26]

Every Joule is Precious: The Case for Revisiting Operating System Design for Energy Efficiency

Amin Vahdat

OXFORD
e-Research
CENTRE

What can we do in math libraries?

- Optimise energy usage rather than performance?
- How much would we give up?
- How many algorithms can be restated minimizing data movement and increasing computation?
- How can we measure “success”?

OXFORD
e-Research
CENTRE

What is needed

- ☐ Support from hardware, low-level systems to provide information on energy usage for operations
- ☐ Tools to provide energy profile for developers
- ☐ Metrics and benchmarks for energy-efficient algorithms/applications

Consistently across platforms!

Profiling for energy

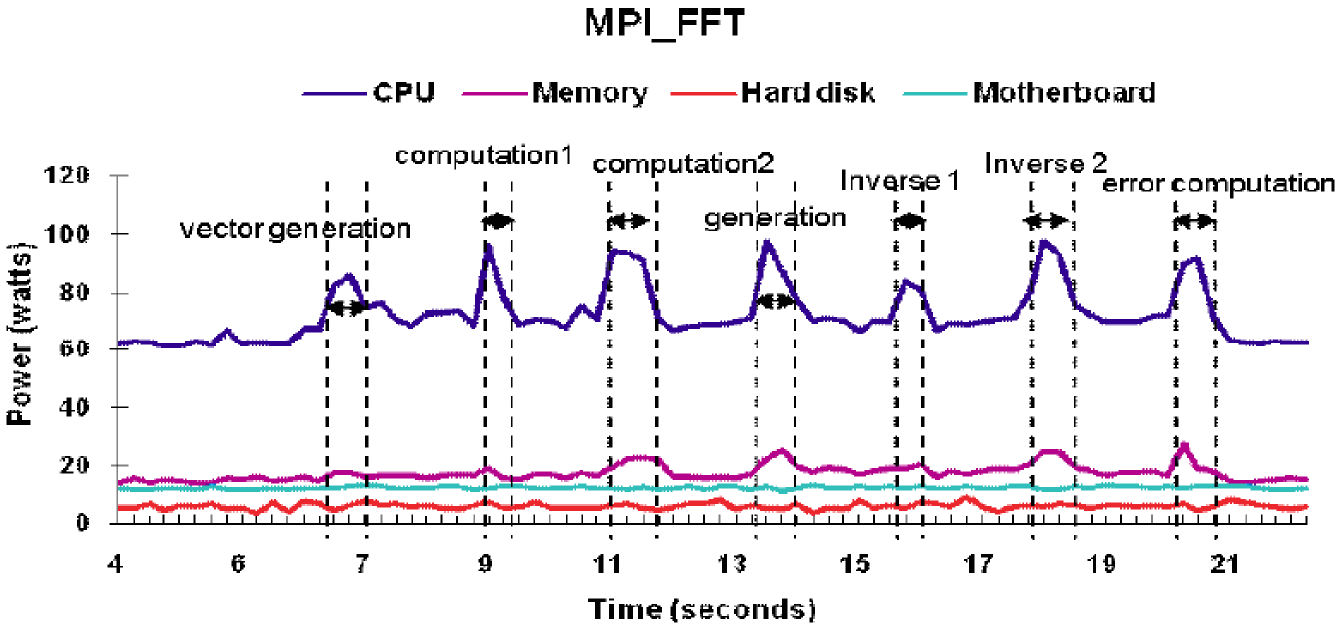


Fig. 5 Detailed power-function mapping of MPI_FFT in HPCC. PowerPack shows processor power rises during computation phases, and drops during communication phases. The seven spikes in processor power profile correspond to vector generation, computation1, computation2, random vector generation, inverse computation1, inverse computation2, and computation of error; the valleys correspond to transpositions that involve inter-processor communications.

Energy Profiling and Analysis of the HPC Challenge Benchmarks

Shuaiwen Song, Rong Ge, Xizhou Feng and **Kirk W Cameron**

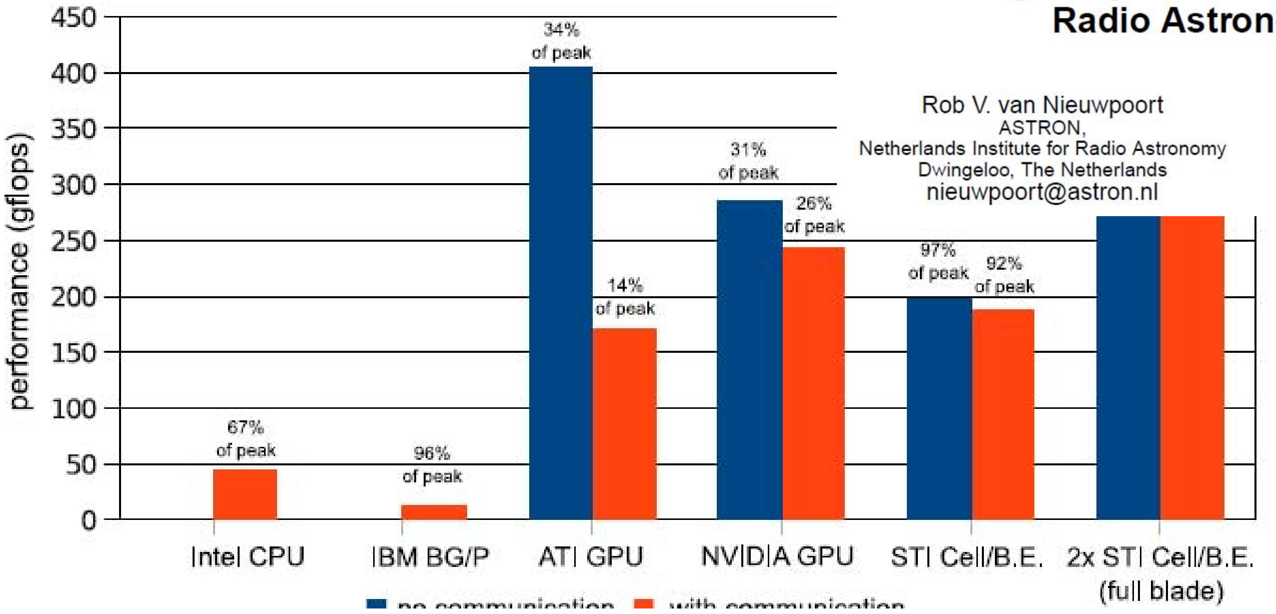
International Journal of High Performance Computing Applications published online 5 June 2009

Experiments in SKA correlation and other components

Using Many-Core Hardware to Correlate Radio Astronomy Signals

Rob V. van Nieuwpoort
ASTRON,
Netherlands Institute for Radio Astronomy
Dwingeloo, The Netherlands
nieuwpoort@astron.nl

John W. Romein
ASTRON,
Netherlands Institute for Radio Astronomy
Dwingeloo, The Netherlands
romein@astron.nl

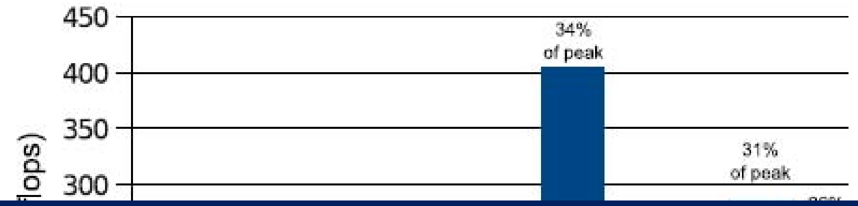


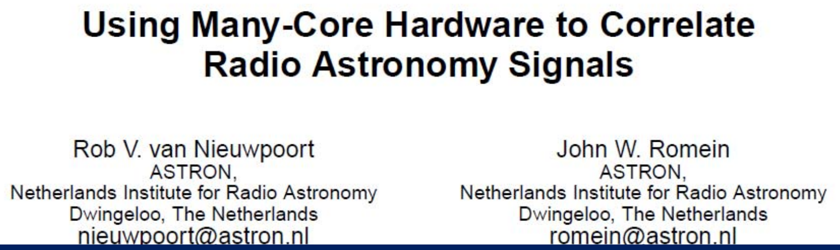
Percentages are th

Figure 4

Architecture	Intel Core i7	IBM BG/P	ATI 4870	NVIDIA Tesla C1060	STI Cell
measured gflops	48.0	13.1	171	243	187
achieved efficiency	67%	96%	14%	26%	92%
measured bandwidth (GB/s)	18.6	6.6	47	94	49.5
bandwidth efficiency	73%	48%	41%	93%	192%
achieved gflops/Watt	**0.37**	**0.54**	**1.07**	**1.00**	**2.67**

Experiments in SKA correlation and other components



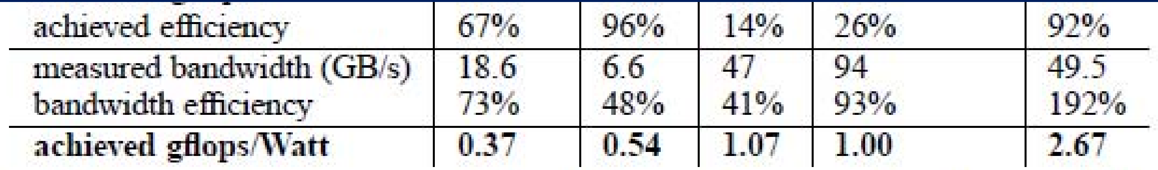

Using Many-Core Hardware to Correlate Radio Astronomy Signals

Rob V. van Nieuwpoort
ASTRON,
Netherlands Institute for Radio Astronomy
Dwingeloo, The Netherlands
nieuwpoort@astron.nl

John W. Romein
ASTRON,
Netherlands Institute for Radio Astronomy
Dwingeloo, The Netherlands
romein@astron.nl

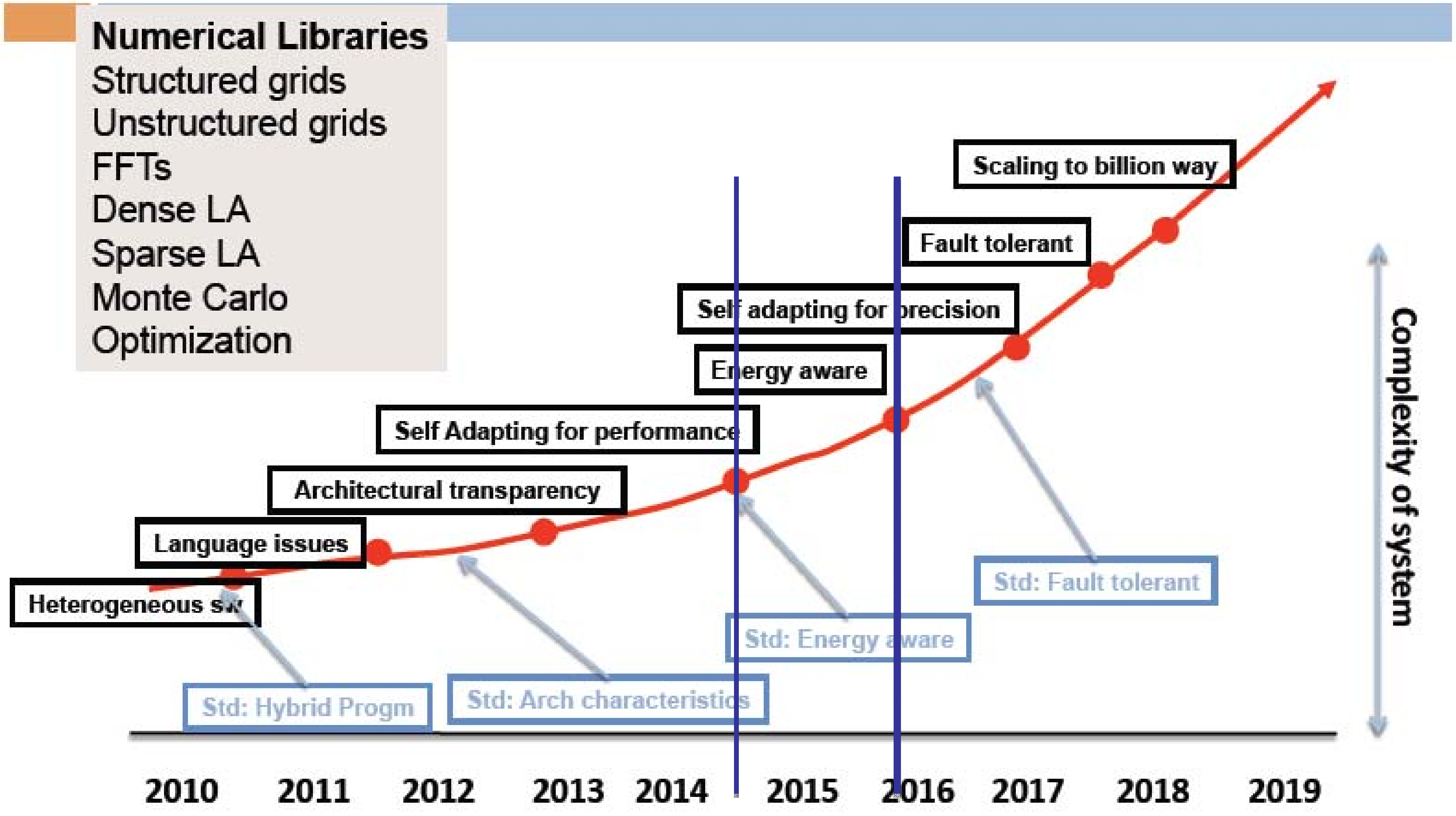
Many individual efforts for specific algorithms

We need a better framework to leverage such efforts



achieved efficiency	67%	96%	14%	26%	92%
measured bandwidth (GB/s)	18.6	6.6	47	94	49.5
bandwidth efficiency	73%	48%	41%	93%	192%
achieved gflops/Watt	**0.37**	**0.54**	**1.07**	**1.00**	**2.67**

4.2.4 Numerical Libraries



Numerical Libraries
- Structured grids
- Unstructured grids
- FFTs
- Dense LA
- Sparse LA
- Monte Carlo
- Optimization

From iesp roadmap 1.0

Conclusions

- Our progress in extreme computing (and to exascale) is constrained by energy consideration (and therefore cost)
- There is a need to enable energy-efficient/energy-aware algorithms across the ecosystem of computing
- Co-design will be essential to allow appropriate architectural and algorithmic decisions to made – application frameworks will help
- Call to action on development of standards/metrics and benchmarks to enable energy measurements and awareness

OXFORD
e-Research
CENTRE

Conclusions

- The heterogeneous platforms lend themselves to energy optimising algorithms.
- We believe appropriate profiling capability will allow developers to create equally efficient performing applications with a lower energy requirement.

There are many other issues that I could have talked about that will cause problems for my SKA colleagues – complexity of the systems and software and usability are notable - I will leave these for next time!

Acknowledgements

My thanks to colleagues on the prepSKA project Steve Rawlings, Aris Karastergiou, Stef Salvini, Ben Mort, Chris Williams, Fred Dulwich and Andy Faulkner. The HPC/NA project was in collaboration with Nick Higham, Iain Duff, Peter Coveney, Mark Hylton and Stef Salvini. I am grateful to Jeyan Thiyagalingam, Simon McIntosh-Smith, Jon Crowcroft and Jaafar Elmirghani for their input and suggestions.

Questions?