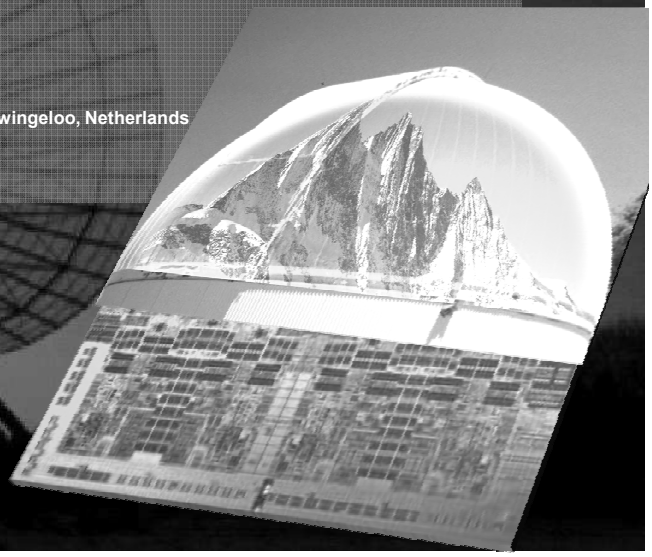
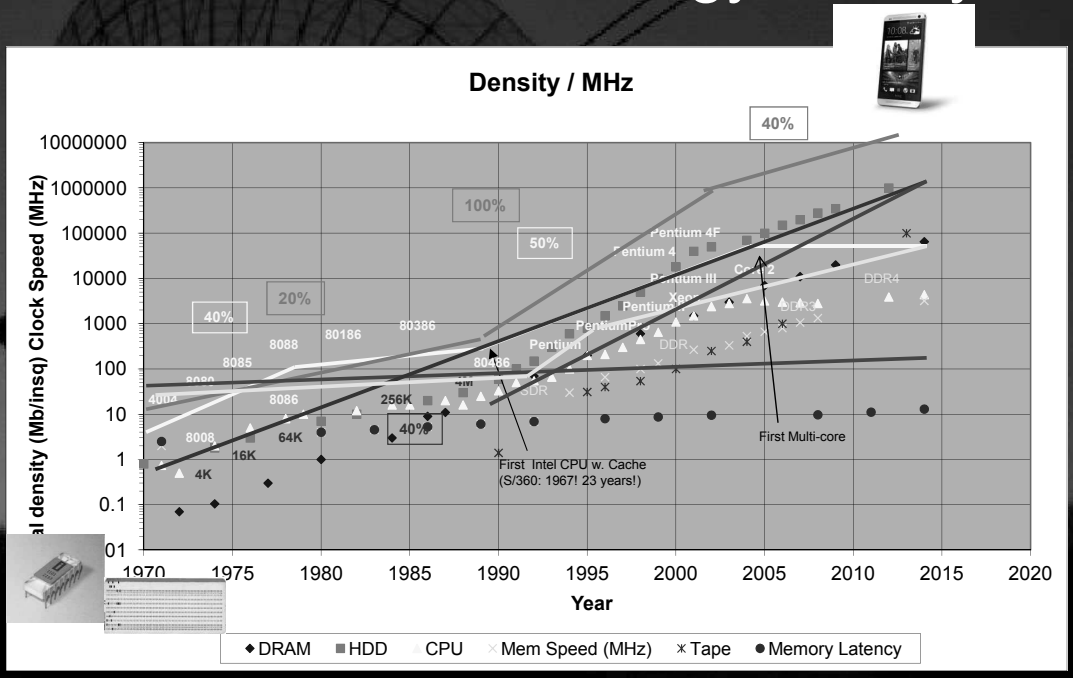
Where do we go now that core performance has stalled?

Dr. Ton Engbersen,
Sci. Dir. ASTRON & IBM Center for Exascale Technology, Dwingeloo, Netherlands
IBM Research Laboratory, Zurich, Switzerland
Member IBM Academy of Technology
apj@zurich.ibm.com



A bit of IT- & Technology History...



ASTRON

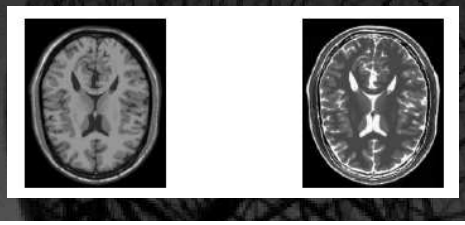
IBM

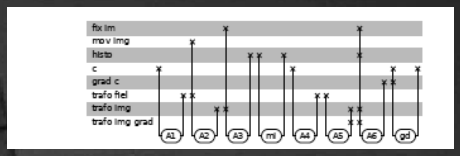
ASTRON

IBM

Parallelism

("the end of software developers paradise")



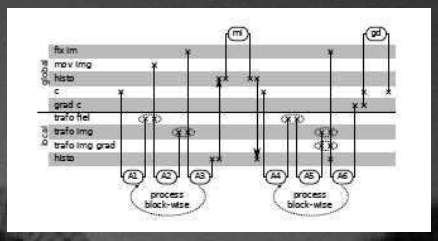


Description	Xeon 3.6 GHz	Core2 3.2 GHz	Cell/B.E.
ITK	62678.5	28698.1	-
Restructuring	3326.3	1067.6	-
Temp buffering	2552.6	813.6	-
Field grouping	2075.5	642.5	-
Gradient grouping	1259.8	407.9	-
Image LUT	1083.4	351.0	1806.3
Manual vectorization	789.9	243.6	669.5
Optimized load	-	-	333.8
Parallelism 1 Chip	-	63.3	43.5
Parallelism 2 Chips	401.4	-	23.14
Double Buffering	-	-	23.11

~500x

~2000x

Tab. 6.1: Runtime per voxel and iteration in nanoseconds depending on optimizations and platform. The Intel Xeon processor was manufactured in 90 nm technology like the Cell/B.E. processor. The Core2 is a more recent x86 processor of the 45 nm generation.



PhD J. Rohrer – 2006-2009
ETH Zürich.

ASTRON

IBM

SKA: What is it?

~0.5M Antennae
.07GHz-0.45GHz.

~0.5M Antennae
.5GHz-1.7GHz.

~3000 Dishes
3GHz-10GHz.

SKA = Square Kilometer Array

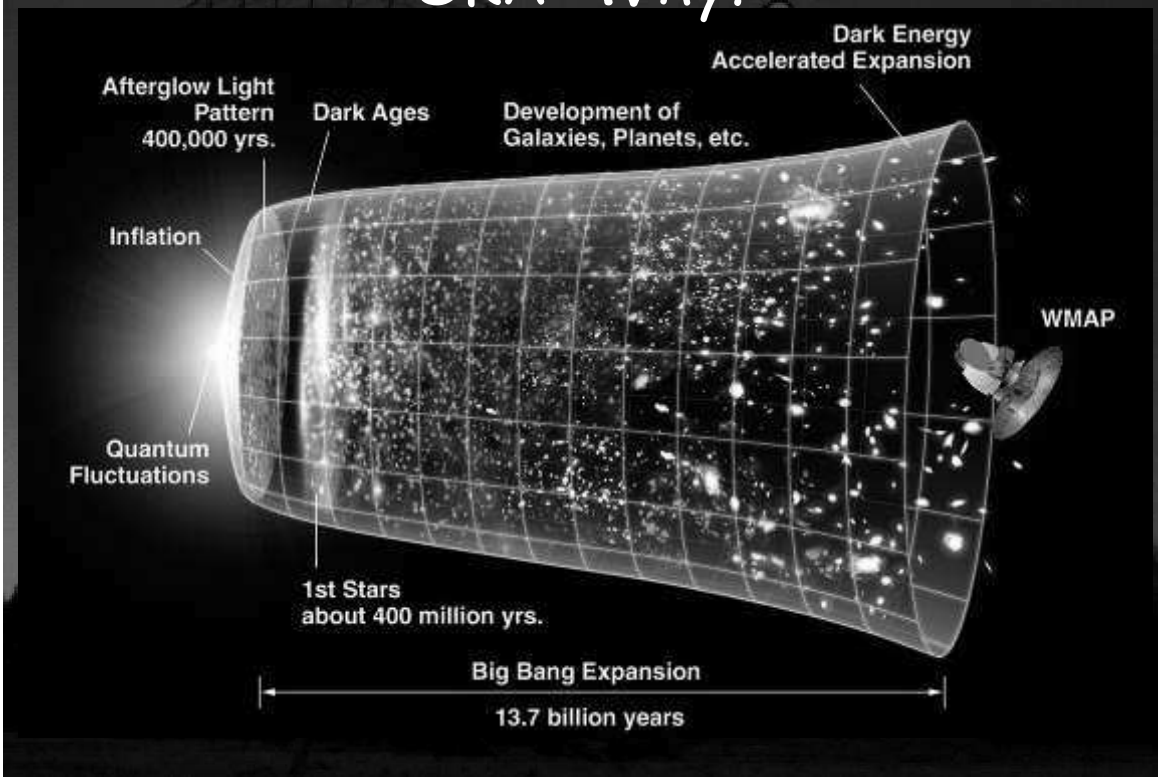
Next generation radio astronomy:
Phase I: Start of deployment 2022
Phase II: Start of deployment 2024

Now:
phase to address fundamental and applied science questions to enable such a system

ASTRON

IBM

SKA: Why?



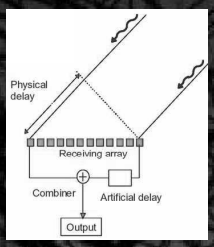
Aperture synthesis

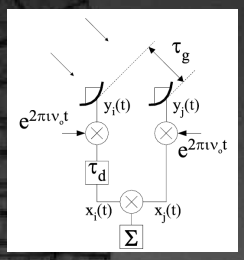
celestial sphere sky sources

beamforming at stations

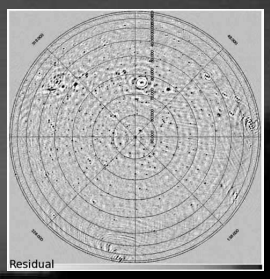
interferometry, correlation of station beams

reconstruction of sky image



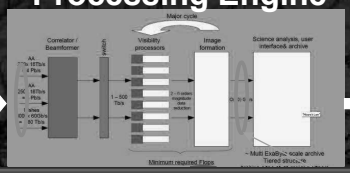


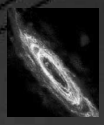
Fourier transform relation $y(t) = \mathbf{w}^H\mathbf{x}(t) \rightarrow \mathbf{R} = E\{\mathbf{y}(t)\,\mathbf{y}(t)^H\}$ inverse Fourier transform



SKA: Data Processing?

Processing Engine





1. 10^9 samples/second * .5M antennae: .5 10^{15} samples/sec.

2. 3.5 10^9 samples/second * .5M antennae: 1.7 10^{15} samples/sec.

3. 2 10^{10} samples/second * 3K antennae: 6.10^{13} samples/sec

Sum = 2 10^{15} samples/second @ 86400 seconds/day:

170 10^{18} (Exa) samples/day. Assume 10-12x reduction @antenna:

14 Exabytes/day (minimum).

Top 500: Sum=123 PFlops. 2GFlops/watt.
➔ 100x Flops of Sum! ➔ ~ 7GWh

➔ **We need the DOME project to improve this!**

ASTRON

IBM

Dome Project:

Research Streams…

Sustainable Computing	Nanophotonics	Data & Streaming

…are mapped to research projects:

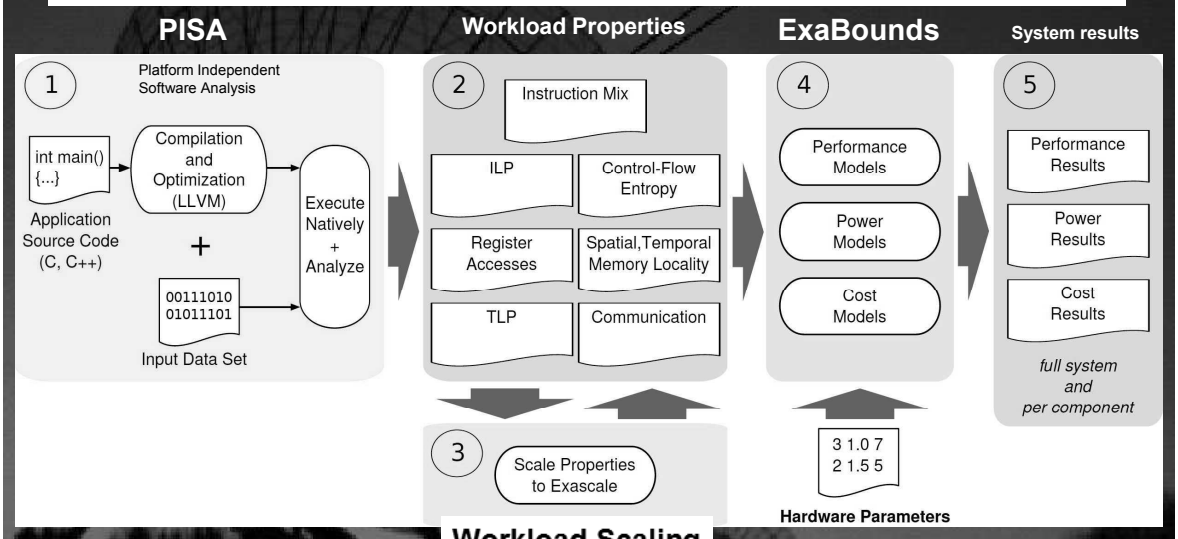
System Analysis		
Algorithms & Machines		
Computing	**Transport**	**Storage**
- Microservers - Accelerators	- Nanophotonics - Real-Time Communications - New Algorithms	- Access Patterns

ASTRON

IBM

Design Space Exploration to Exascale

Goal: Create a holistic design-space exploration tool to overcome fundamental technology limits in data centers, servers and exascale systems by use of a novel formal method that captures first principles in form of equations compounded with boundary conditions (power, required throughput, I/O, technology parameters, architecture).



Results:

- Improved accuracy 1-2 orders of magnitude (vs. traditional linear regression)
- Design Space Exploration of 324 architecture design points in ~50 min. on laptop

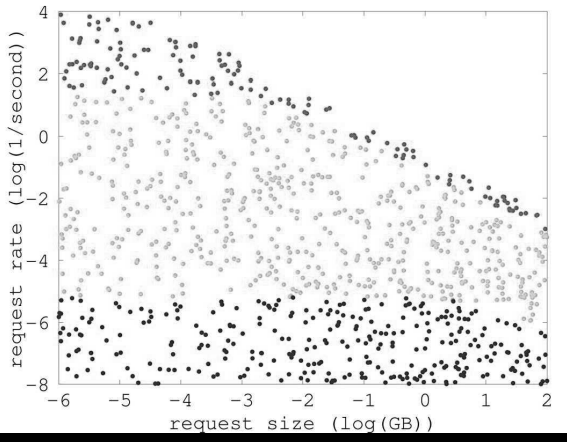
A. Anghel, L. Vasilescu, R. Jongerius, G. Dittmann, G. Mariani, "**An Instrumentation Approach for Hardware-Agnostic Software Characterization**", G. Mariani, A. Anghel, R. Jongerius, G. Dittmann, "**Scaling Application Properties to Exascale**", Both: ACM Computing Frontiers, May, 18-21, 2015, Ischia, Italy.

Storage tiers:

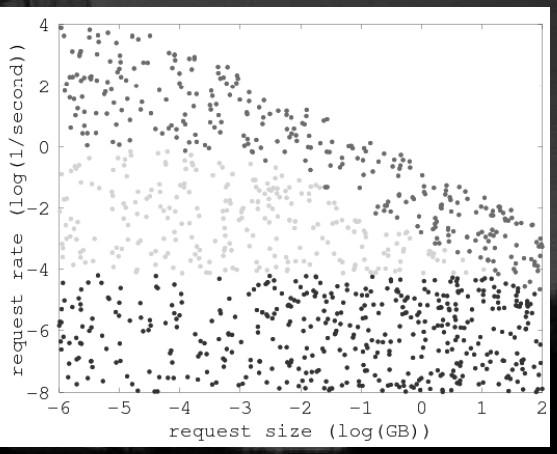
Data units assignment example (1000 100GB chunks)

- Red: SSD
- Green: HDD
- Blue: Tape

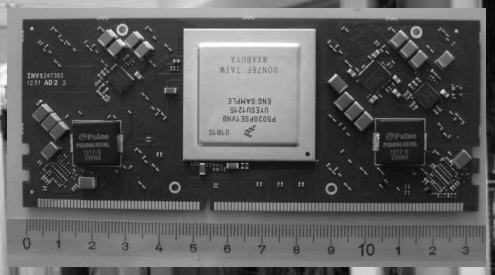
Budget = $23,000
Mean Response = 0.27 sec

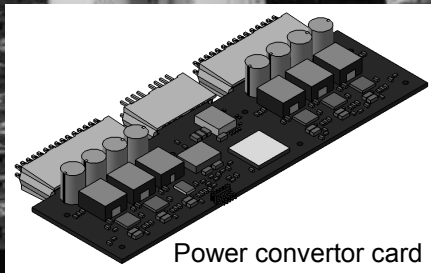


Budget = $35,000
Mean Response = 0.0017 sec



Micro Servers:



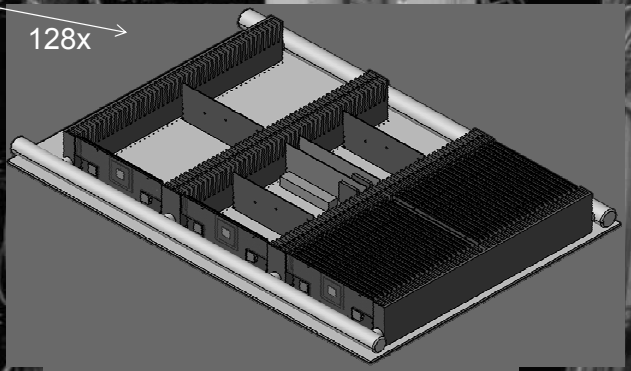


Power convertor card

µServer:

The integration of an entire server node motherboard* into a *single microchip* except DRAM, NOR-boot flash and power conversion logic.

128x



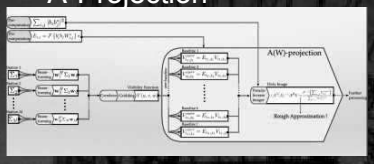
DOME Microserver mechanical drawing of 2U node:
512 x 64-bit POWER Cores @ 2.2GHz, 2TB DRAM

R.P. Luijten, R. Clauberg, A. Doering, D. Pham, A. Nguyen, M. Pandya, **"Energy efficient MicroServer based on 1.8GHz 12 core 185K Coremark 28nm 64-bit SoC for big data applications featuring 159GB/s/liter memory BW system density"**, IEEE International Solid-State Circuits Conference (ISSCC), Feb. 22-26, 2015, San Francisco, USA.

Basic Algorithms:

Basic radio-interferometry:

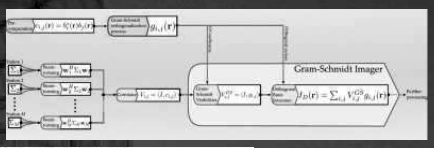
A-Projection

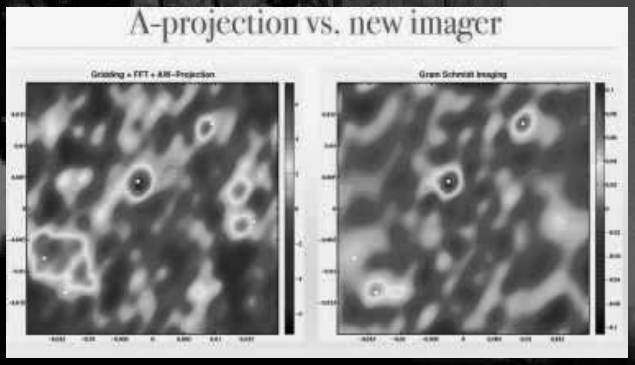


A new imaging domain:

More natural domain,

- No gridding
- No FFT
- More flexible





Publication in preparation

Real-Time Communication:

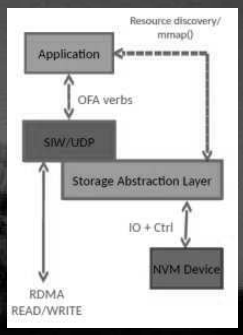
- **Power consumption of data transfer**
- **Efficient data transfer – fast, zero-copy**
- **"Big Data" (sensor) processing**

<u>Unreliable RDMA protocol, with direct access to FLASH:</u>

- **Soft-iWARP /UDP RDMA: 30% improved latency over TCP,**
- **~ 50% less power / Gb/s then TCP/IP data transfer.**

"FlashNet":

- Extensions to S-iWarp to access remote flash
- Full RDMA end-to-end semantics
- Mediate between varying Flash access delay and network
- Stall/resume data transfer upon media availability
- Request I/O page before IO mem. Operation
- Sync. After read/write completion




ASTRON
IBM
Take away:
In the Exa-data-regime:
1. Design exploration becomes key to control power dissipation
2. Data access&movement is the key problem (not computation)
Consequently "Mind-your-DATA-access&movement"
3. Re-thinking of the algorithm to do the task is a good starting point