The DOME Project

Prof. Dr. Ton Engbersen,
Prof. Data Science Engineering – Rijksuniversiteit Groningen, NL
Sci. Dir. ASTRON & IBM Center for Exascale Technology, Dwingeloo, Netherlands
IBM Research Laboratory – Zurich, Switzerland
Member IBM Academy of Technology
apj@zurich.ibm.com
The World is Our Lab

World's largest information technology research organization

More than 3,000 scientists and engineers

IBM spent $6.2B on R&D in 2014

- Austin
- T.J. Watson
- Almaden
- Brazil
- Ireland
- Haifa
- Africa
- India
- China
- Tokyo
- Australia
- Zurich
- IBM spent $6.2B on R&D in 2014

The World is Our Lab
**SKA: What is it?**

- **Top 500: Sum=123 PFlops.**
- **2GFlops/watt.**
- **≈ 7GWh**

1. $10^9$ samples/second * .5M antennae: $5 \times 10^{15}$ samples/sec.
2. $3.5 \times 10^9$ samples/second * .5M antennae: $1.7 \times 10^{16}$ samples/sec.
3. $2 \times 10^{10}$ samples/second * 3K antennae: $6.10^{13}$ samples/sec

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

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

14 Exabytes/day (minimum).
SKA: Processing?

Too hard

Moore’s law

Too easy *(for us)*
SKA: Processing?

There is only one way to get here:

Smart methods & algorithms

Fast hardware
SKA: Data Moving?

<table>
<thead>
<tr>
<th>TByte</th>
<th>PByte</th>
<th>EByte</th>
</tr>
</thead>
<tbody>
<tr>
<td>100Mb/s</td>
<td>~1 Day</td>
<td>~2.5 Years</td>
</tr>
<tr>
<td>1 Gb/s</td>
<td>~2 Hours</td>
<td>~3 Months</td>
</tr>
<tr>
<td>10 Gb/s</td>
<td>~10 Min.</td>
<td>~1 Week</td>
</tr>
<tr>
<td>100 Gb/s</td>
<td>~1 Min.</td>
<td>~16 Hours</td>
</tr>
</tbody>
</table>
Data Gravity!

Snowball

Into cloud (50TB): 0.00$
Out of cloud (50TB): 1500.00$

(Mar. 1st, 2016)
Micro Cloud brings the benefits of cloud computing to data that is difficult to move.

Micro Cloud consists of 3 components:
- A self-managing on-premise appliance
- Traditional cloud for SaaS
- APIs to move SaaS to appliance

Micro cloud downloads computation from a cloud site to the appliance
- Computation executed in a safe and controlled environment within the appliance

Specific Case: A Bank
- Operates in many countries, data cannot move off-premises due to compliance requirements
- Micro cloud brings analytics in cloud to on-premise data, enables comparative analytics
Dome Project:

Research Streams...

- Sustainable (Green) Computing
- Nanophotonics
- Data & Streaming

...are mapped to research projects:

- System Analysis
  - Algorithms & Machines

- Computing
  - Microservers
  - Accelerators

- Transport
  - Nanophotonics
  - Real-Time Communications
  - New Algorithms

- Storage
  - Access Patterns

...plus an open user platform:

- Student projects
- Events
- Research Collaboration

33M€ 5-year Research Project: 76 IBM PY (32 in NL); 50 ASTRON PY
Aperture synthesis

Beamforming at stations

Interferometry, correlation of station beams

Reconstruction of sky image

Central Signal Processor (CSP)

Science Data Processor (SDP)

Archive

Algorithms and Machines (P1)

Access Patterns (P2)

Nanophotonics (P3)

Microservers (P4)

Accelerators (P5)

New Algorithms (P6)

Real-Time Communications (P7)
Aperture synthesis

Beamforming at stations

Interferometry, correlation of station beams

Reconstruction of sky image

Central Signal Processor (CSP)

Science Data Processor (SDP)

Archive

Algorithms and Machines (P1)

Access Patterns (P2)

Nanophotonics (P3)

Microservers (P4)

Accelerators (P5)

New Algorithms (P6)

Real-Time Communications (P7)
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).
Aperture synthesis

Beamforming at stations

Interferometry, correlation of station beams

Reconstruction of sky image

Central Signal Processor (CSP)

Science Data Processor (SDP)

Archive

Algorithms and Machines (P1)

Access Patterns (P2)

Nanophotonics (P3)

Microservers (P4)

Accelerators (P5)

New Algorithms (P6)

Real-Time Communications (P7)
"cognitive" storage

Interactions with SKA SDP.DATA on tape cost, IO-tracing tool....
Cognitive storage System - Design

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
Aperture synthesis

Beamforming at stations

Interferometry, correlation of station beams

Reconstruction of sky image

Central Signal Processor (CSP)

Science Data Processor (SDP)

Archive

Algorithms and Machines (P1)

Access Patterns (P2)

Nanophotonics (P3)

Microservers (P4)

Accelerators (P5)

New Algorithms (P6)

Real-Time Communications (P7)
Synergy between IBM projects and the SKA system:
Parallel optical interconnects → Analog Fiber Links
WDM and single mode silicon photonics → Photonics Beamforming

Contributions
• Antenna to base station electrical – optical – electrical link analysis & demo
• Optical beam forming using silicon photonics
Conclusions on a RFOF optical link:
- Established generic analog optical link model
- Multimode link realized
  - Can reach 7 km, possibly also 10 km
  - Limiting factor Receiver Amplifier (TIA)
  - Best improvement through better PD and higher VCSEL Slope Efficiency
- Actual implementation with single mode fiber
  - Fiber cost dominates

Recently an experimental system (single mode fiber) has been tested in Australia
Beamforming = Reducing Data:  
- Electronic/Digital Beamforming «in» the dish 
  risks to interfere with signal reception (RFI)  
- Optical Beamforming ... does NOT!!

Photonics Beamformer Requirements: 
- Frequency Range: ~3 GHz
- Array Dimensions: 10 x 11 Elements
- Element Spacing: 21 mm
- Array Size: 22.4 x 22.4 cm
- Beam-Steering: +/- 30°

Currently being measured
μServer: The integration of an entire server node motherboard* into a *single microchip* except DRAM, Nor-boot flash and power conversion logic.

*This does NOT imply low performance!
Microserver – Built-up

- **Compute Node FR4 board**
  - 3 layer Laminated Copper
  - 12V GND GND 12V
  - 7.6 mm

- **Power converter boards**
  - Storage boards

- **Water**
  - Compute Nodes

- **40% less energy compared to conventional systems**
  - 90% of waste heat can be reused
Planned System: 2U rack unit

40% more performance @ 70% of node level energy consumption
→ 2x more operations per Watt

2015 IEEE International Solid-State Circuits Conference
Aperture synthesis

Beamforming at stations

Interferometry, correlation of station beams

Reconstruction of sky image

Central Signal Processor (CSP)

Science Data Processor (SDP)

Archive

Algorithms and Machines (P1)

Access Patterns (P2)

Nanophotonics (P3)

Microservers (P4)

Accelerators (P5)

New Algorithms (P6)

Real-Time Communications (P7)
Image-Domain gridding

Explore inherent parallelism in gridding...?

Convolution in Fourier Domain = multiplication in image domain

- ~32 x 32 subgrids
- Local memory
- parallelism

It works!

Status:
- On CPU (!) 25x faster then LOFAR (CPU) imager
- Presented at GPU-Technology Conf. 2015
Aperture synthesis

Beamforming at stations

Interferometry, correlation of station beams

Reconstruction of sky image

Central Signal Processor (CSP)

Science Data Processor (SDP)

Archive

Algorithms and Machines (P1)

Access Patterns (P2)

Nanophotonics (P3)

Microservers (P4)

Accelerators (P5)

New Algorithms (P6)

Real-Time Communications (P7)
Aperture synthesis

Beamforming at stations

Interferometry, correlation of station beams

Reconstruction of sky image

Central Signal Processor (CSP)

Science Data Processor (SDP)

Archive

Algorithms and Machines (P1)

Access Patterns (P2)

Nanophotonics (P3)

Microservers (P4)

Accelerators (P5)

New Algorithms (P6)

Real-Time Communications (P7)
Efficiency of data transport

1. Power consumption of data transfer and processing
2. Efficient data transfer to data processing entity
3. Intermediate ingress data buffering at proc. entity
RDMA over UDP vs. TCP/IP

- RDMA over TCP/IP can lose up to 35% of the peak IOPS performance, whereas RDMA over UDP delivers within 85-90% of the peak IOPS performance.
- With an offloaded RDMA-stack access latencies improve by 15%.
- It is expected that a large part of SDP will work with FLASH buffers for efficiency and speed. (SDP)

**FlashNet**
- Extension to SIW to access remote flash storage
- Flash storage is managed efficiently
- Full RDMA end-to-end semantics

**Extensions to host RDMA protocol processing**
- Mediation between varying Flash access delay and network
- Stall/resume data processing upon media availability

**Extensions to storage abstraction layer (SAL)**
- Request I/O page before Read/Write IO memory operation
- Synchronize after Read/Write completion
Aperture synthesis

Beamforming at stations

Interferometry, correlation of station beams

Reconstruction of sky image

Central Signal Processor (CSP)

Science Data Processor (SDP)

Archive

Algorithms and Machines (P1)

Access Patterns (P2)

Nanophotonics (P3)

Microservers (P4)

Accelerators (P5)

New Algorithms (P6)

Real-Time Communications (P7)
Conclusion

1. Very active workstreams with focus on SKA, may-be mainly post SKA-1, but many with (partial?) SKA-1 opportunities.

120+ Scientific (peer reviewed) publications since start, many more accepted & in preparation, P1, P2, P3, P4, P5, P6, P7

2. Significant impact on current SKA-consortia thinking:

P1 – being used in SDP consortium, ‘chip’ approach for SKA-2, MFAA
P2 – being used in sizing Regional SKA Science Data centers
P3 – Optical transmission in test in Australia site, optical beamformers SKA-2
P4 – Microserver evaluation for SDP (and other applications)
P5 – GPU programming already today improving LOFAR data processing
P6 – Mathematical re-thinking of radio-interferometry fundamentals (next talk)
P7 -- RDMA over UDP & FLASH - Plan of record SDP-ingest.