Welcome

Optics for the Cloud – How Microsoft Research are re-thinking future Data Centre technology

Dr Paolo Costa
Microsoft Research Researcher

Dr Scarlet Schwiderski-Grosche
Microsoft Research Principal Research Program Manager

Dr Ariel Gomez Diaz
Microsoft Research Researcher
Azure Infrastructure

Data centres deployed in 50+ regions, 10,000s of kilometres of fibre

10-20 data centres / region

≤ 60 km

~100,000s servers / data centre
How do we scale cloud capacity?

Faster servers

- Mem
- CPU
- Disks
- FPGA
- GPU
Data doubling every two years

Need for low latency at scale

ML and AI driving this further

Source: IDC 2014
How big is one Zettabyte?

• 1 ZB = $10^{21}$ bytes = 1 Billion TB

• 10 TB HDD

• Need 100 Million of those

• That’s a 2500 km high tower

• Data also needs to be transmitted across the network

• 100 Gbps data centre network link

• That’s 2500 years for transmission
Driving exponential growth
Pushing compute growth

Hyperscale cloud hardware design in collaboration with the Open Compute Project (OCP)
Pushing storage growth

Custom storage racks optimised for cold storage

Project Pelican
Pushing networking growth

Custom data centre interconnect technology at multi-terabit speeds

Project Madison
Pushing networking growth
Driving exponential growth

- Compute (FLOPS/$)
- Storage (GB/$)
- Network (Gbps/$)
Disruption

- FLOPS/$
- GB/$
- Gbps/$

Compute | Storage | Network
Disruption ... and opportunities

- FLOPS/$ for Compute
- GB/$ for Storage
- Gbps/$ for Network
Impressive advancements in optical technologies

Cutting-edge technologies in academic labs across the world

Southampton University, University College London, Cambridge University, Eindhoven University of Technology (Tu/e) and University of California, Santa Barbara (UCSB)
A big opportunity ... cloud-first approach
Optics for the Cloud

Application

Network

Storage

Physical

New scenarios

New hardware/software

New media

New optics

Innovation across the cloud stack

www.opticsforthecloud.com
Data Center Networks Today

Racks of servers interconnected by electrical packet switches

1) Optical-to-electrical conversion

2) Address lookup

3) Packets are queued on output port

4) Electrical-to-optical conversion

~100,000 servers
Moore’s Law for Networking

**Electrical switching**

Double the speed every two years, *at the same cost*
Free switch scaling enabled keeping network cost low and constant
Disruption

- Electrical switching
- Optical switching

- Cloud network cost
Why Optics?

1. Already used for transmission
2. Non-CMOS based
How can we switch packets using light?

1) Arrayed Waveguide Grating Router (AWGR): passive component (no transistors or buffers)

2) Wavelength-based forwarding

3) Fast tunable laser
How fast can we tune the laser?

Utilization < 30%

20 ns 20 ns 100 ns 20 ns

Off-the-shelf laser + custom algorithms

256-byte packet @ 100 Gbps

Need for a fundamentally different approach
Microsoft Custom Tunable Laser

PIC: Photonic Integrated Circuit (Indium-phosphide)
Fast Tunable Laser

- Tuning time:
  - Off-the-shelf laser: ~1ms (10,000x faster)
  - MS custom laser: <1ns (tuning time ~10x faster)

- Utilization: <30%

- 256-byte packet at 100 Gbps (100x faster)

- 20 ns + custom algorithms
All-optical Data Center Network

Speed-of-light networking
No packet processing or buffering within the network core
Bufferless: blessing or curse?
The lack of buffer requires tight synchronization among nodes
Developed a fully decentralized solution with nanosecond accuracy
Clocks are not synchronized

Synchronization accuracy = 100s picoseconds
(less than 1 billionth of a second)
Project Timeline

- 2016: <1ns laser tuning, Nanosecond time synchronization
- 2018: Scaling, Deployment hurdles, Integration, Reliability
- 2022+: Production
Single pool of resources

Data Center As a Computer
Which new application scenarios can be enabled?
Project Silica
Data volume is growing...

Data generated per year

Volume of data

Years
Data volume is growing... but storage is expensive!

- Data generated per year
- Data we can afford to store
Data volume is growing... but storage is expensive!
Today’s archival storage technologies

<table>
<thead>
<tr>
<th>Media Lifetime [Years]</th>
<th>Total Cost ($ / GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Short</td>
<td></td>
</tr>
<tr>
<td>Long</td>
<td></td>
</tr>
</tbody>
</table>
Today’s archival storage technologies

- Tape: ~10 TB / cartridge
- HDD: ~14 TB / disk

<table>
<thead>
<tr>
<th>Media</th>
<th>Short</th>
<th>Low</th>
<th>Total Cost ($ / GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tape</td>
<td>~10 TB</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>HDD</td>
<td>~14 TB</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>
Today’s archival storage technologies

<table>
<thead>
<tr>
<th>Media</th>
<th>Lifetime [Years]</th>
<th>Total Cost ($ / GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tape</td>
<td>~10 TB / cartridge</td>
<td>Low</td>
</tr>
<tr>
<td>Optical disk</td>
<td>300 GB</td>
<td>Low</td>
</tr>
<tr>
<td>Magnetic media</td>
<td>HDD ~14 TB / disk</td>
<td>High</td>
</tr>
</tbody>
</table>

Short

Long
Today’s archival storage technologies

- **Silica**
  - Up to 360 TB
- **Optical disk**
  - 300 GB
- **Tape**
  - ~10 TB / cartridge
- **HDD**
  - ~14 TB / disk

<table>
<thead>
<tr>
<th>Media</th>
<th>Lifetime [Years]</th>
<th>Total Cost ($ / GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnetic media</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>Long</td>
<td>High</td>
</tr>
</tbody>
</table>

- **Low**
- **High**
Archival storage in Glass

- Raw material is SiO2 (Quartz)
- 10000s year lifetime
- Write Once Read Many (WORM)
- Persistent: no bit rot or media decay
- Resilient: EMF-proof, resistant to heat.
Writing data into the glass...

Pulsed lasers →

Femtosecond (fs) pulsed laser

\[ 1 \text{ fs} = 10^{-15} \text{ seconds} \]

0.0000000000000001 seconds
Picosecond VS femtosecond laser writing

Picosecond \((10 \times 10^{-12} \text{ s})\) laser induces voids \textit{with external stress}\n
Femtosecond \((10 \times 10^{-15} \text{ s})\) laser induced \textit{small} gratings in \textit{quartz glass}.\n
![Image of picosecond laser writing](image1.png)

![Image of femtosecond laser writing](image2.png)
Femto-second laser writing

- Laser beam is focused inside the glass
- At the focus, the glass is modified creating a ‘voxel’
Voxel structure: nanograting

- Multi-level encoding using shape and orientation of structure

Self-Organized Nanogratings in Glass Irradiated by Ultrashort Light Pulses
Yasuhiko Shimotsuma, Peter G. Kazansky, Jiarong Qiu, and Kazuoki Hirao
Disaggregated Write-Read technology

Write Head

Glass library

Read Head
Reading system

- Source is an LED
- Optical imaging setup
- Parallel readout
Polarization readout

- Several images per layer are acquired.
- Voxel traces are decoded
- Multiple bits of information per voxel
Read head demo:

Glass slide with many tracks ('dots' to the naked eye)

Each 'dot' is a 3D array of voxels
Reading head demo:

Stacks of voxels

XY-Scan...
Optics Research inspired and designed for the Cloud

http://opticsforthecloud.com/

Thank You
Conclusions
Things to do next

Session Feedback
Please rate this session in the Future Decoded app!

Microsoft UK AI Research Report
Download the AI Report at http://aka.ms/UKAIreport

Visit our Hands-on Labs on Level 3
Try technology out with on-demand labs and expert help

Go deep with Documentation
http://docs.microsoft.com
Microsoft Ignite | The Tour
London
ExCeL London
February 26-27, 2019

Learn new ways to code, optimise your cloud infrastructure, and modernise your organisation with deep technical training.

Register for free on Wednesday, October 31, 2018.

aka.ms/IgniteTourLND
Thank You