HP and Exascale

HP and exascale

– HP is committed to pursue 100 peta and exascale systems for HPC, leveraging new technologies

– ESSN (Enterprise Servers, Storage, and Networking) business

  - For 3Q2010 $4.4B in revenue (14% of HP's revenue)

– HP Labs – Intelligent Infrastructure Lab (Norm Jouppi)

  - Processors, memory
  - Photonics
  - Systems software, programming models
  - Sensor networks
  - Many academic research partners under IRP (www.hpl.hp.com/open_innovation/), and summer internships.

– Other key labs: Networking, Memristors, Sustainability

Technology transfer

– A tradition: HP Labs ←→ ESSN

– Power and cooling
    - Top issues for commercial customers
    - Key issue for effective exascale

– Massive scale-out is a large vertical market
    - Photonics is an enabler for the fabric
    - Photonics is under consideration for future servers

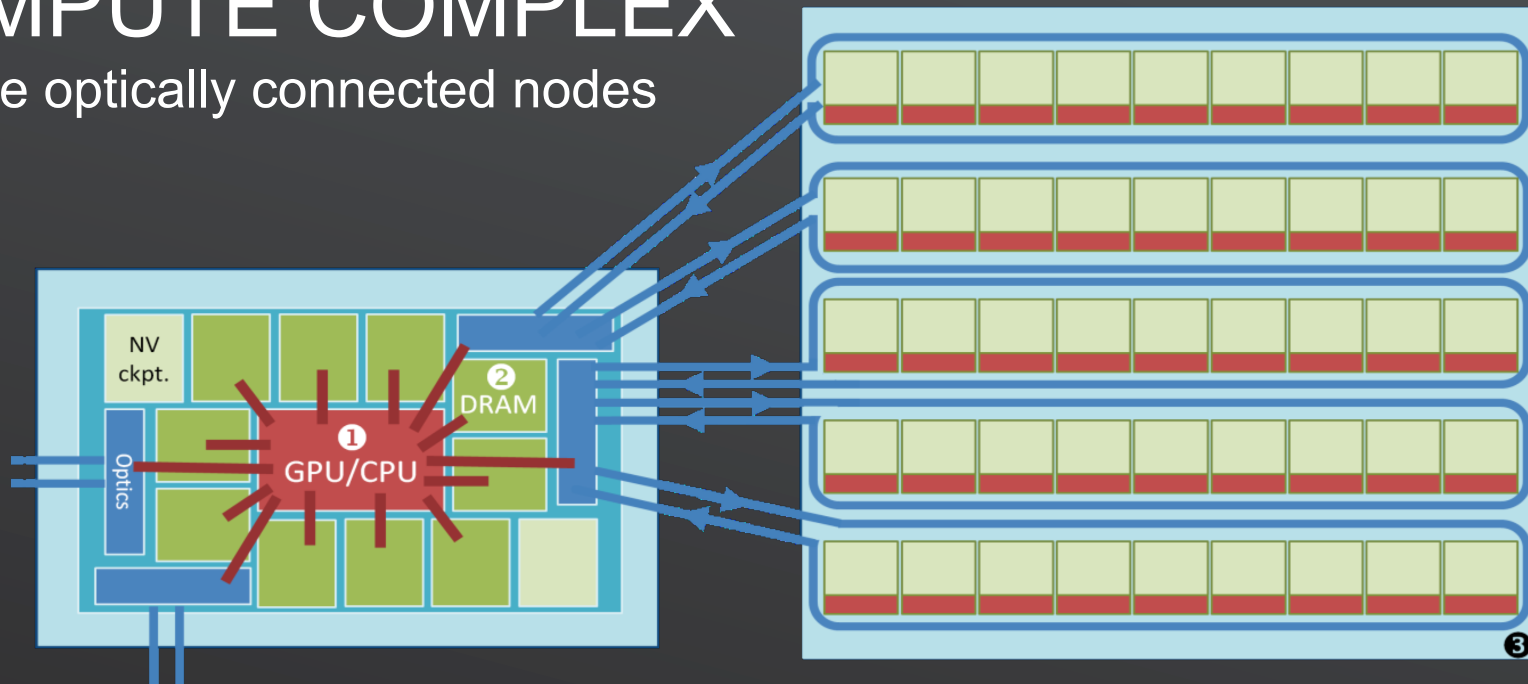
– Networks, photonics

– Software
  - System management
  - Networking
  - Resilience, programming

– New processor/core compute strategies
  - ARM, GPU

COMPUTE COMPLEX

Flexible optically connected nodes

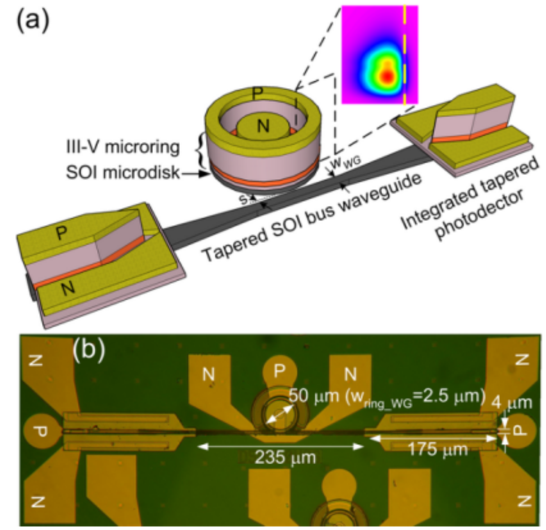


– Single-chip, integrated GPU/CPU (1)
– 3D-stacked DRAM cubes through high speed low-energy links (2)
– Silicon interposer provides connectivity to tightly couple the GPU/CPU chip, DRAM cubes, checkpoint stores, and optical interface dies.
– Optically connected nanostores of NVM (3) for large memory

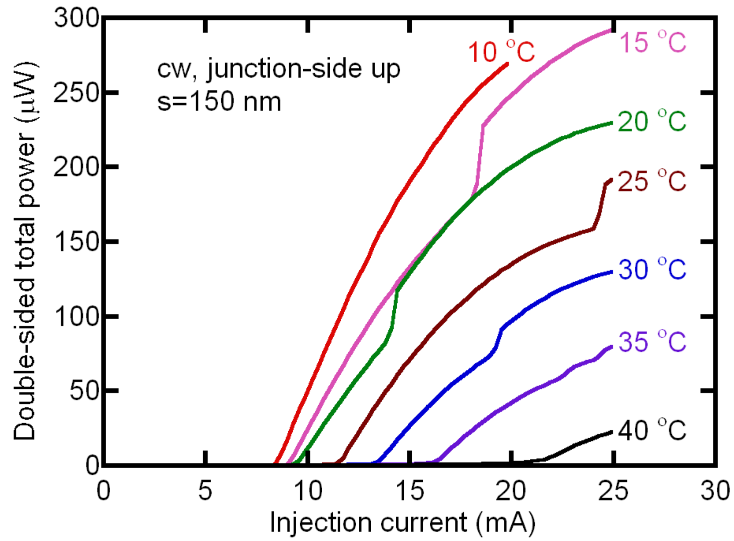
Photonics

HYBRID RING LASER

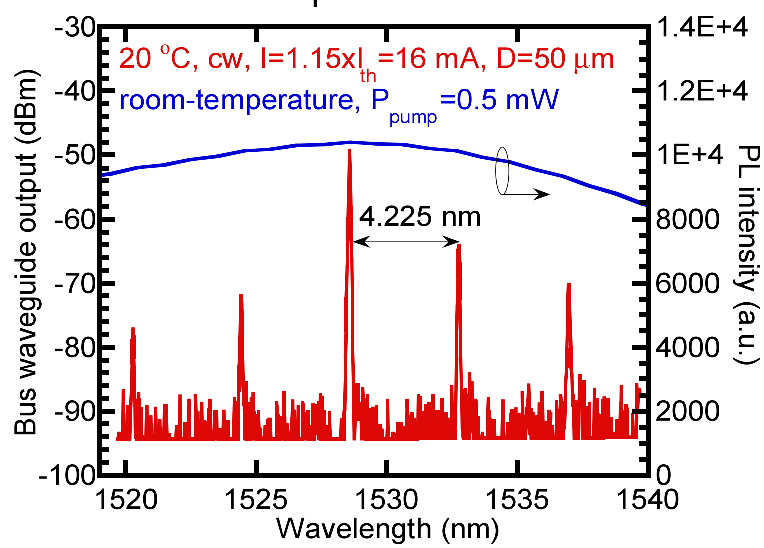
- Hybrid Si-InAlGaAs platform
- Wafer bonding
- Self-aligned process
- 300 µW output
- Up to 65°C lasing observed
- Single transverse mode
- 10 GHz 3 dB BW expected
- Can be used for CWDM link



LI curves



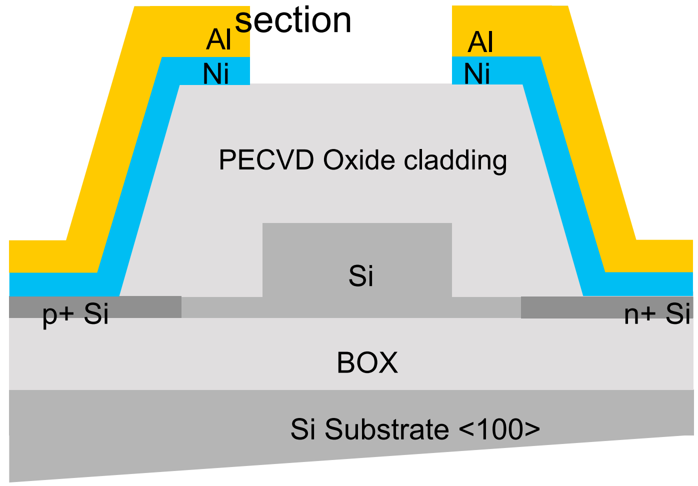
Spectrum



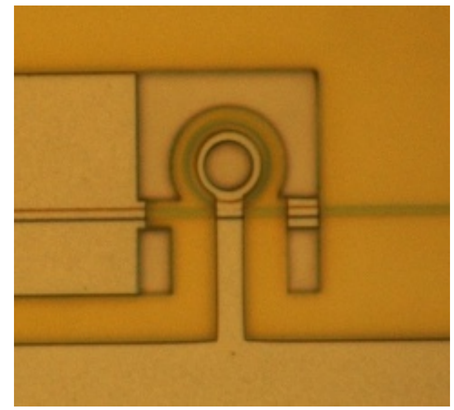
MICRO RING MODULATORS

- 10 µm silicon ring resonators
  - Charge injection
  - 1310 nm (compatible with Ge detectors)
- Results
  - Q > 10,000
  - 700 Ω series resistance (pads)
  - 0.18 nm shift (thermal limit)
  - 18 dB extinction
  - 45 fJ/bit
  - 6 Gbps modulation (with pre-emphasis)

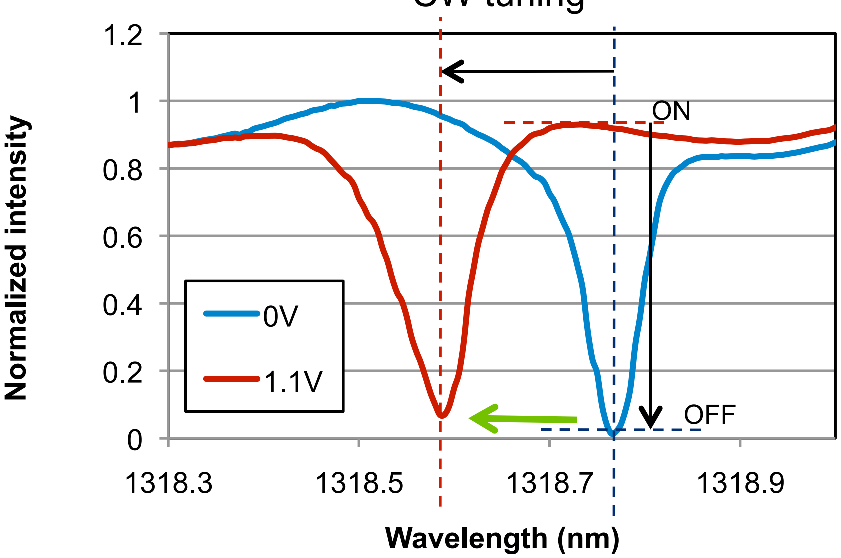
Cross section



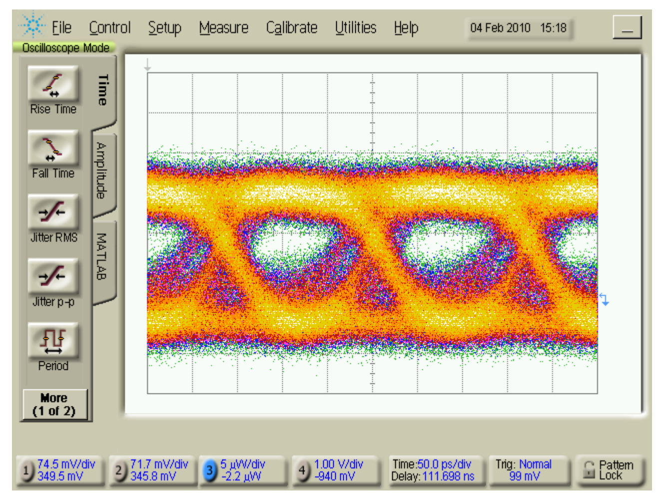
Top view



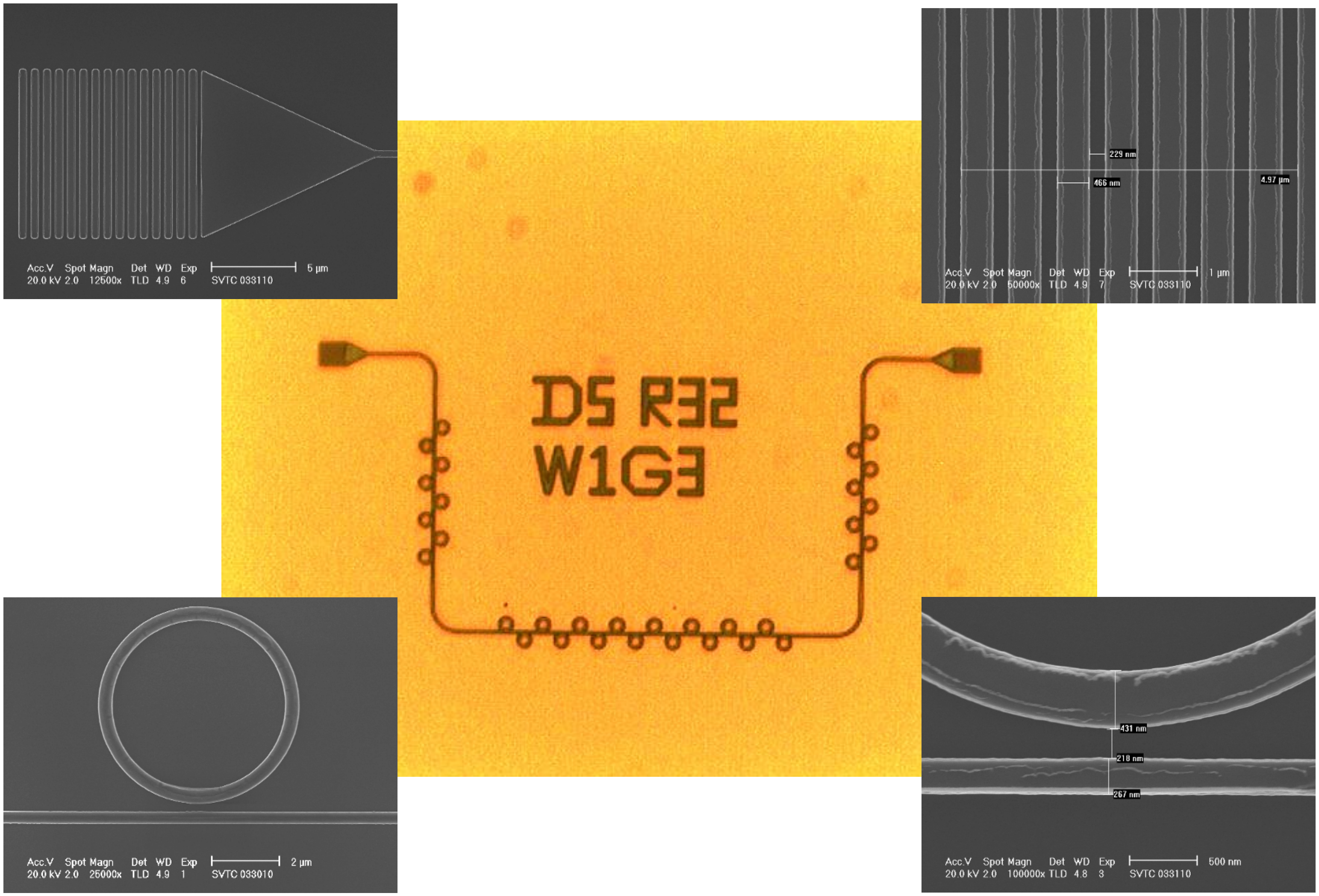
CW tuning



Eye diagram NRZ 6 Gbps



LARGE-SCALE CMOS FABRICATION



Interconnect

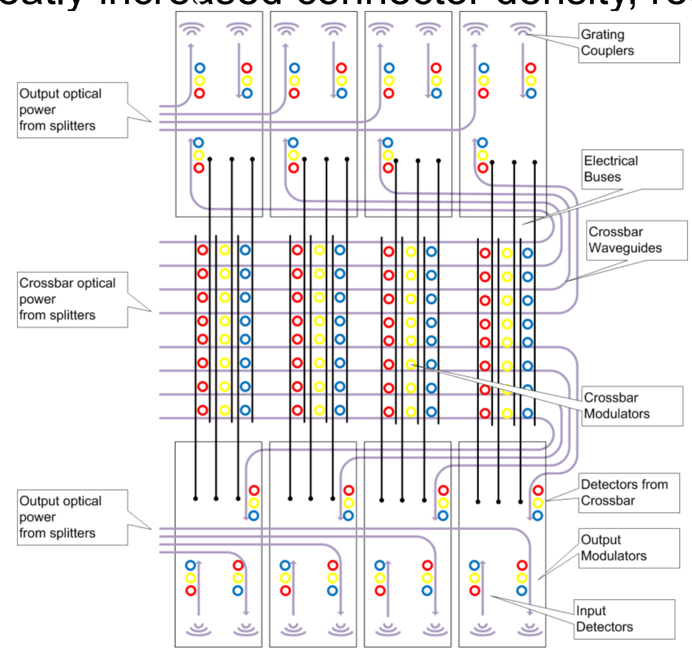
HIGH RADIX PHOTONIC SWITCH

ELECTRONIC

– Switch port count limited by pin count & IO power
– Additional external transceivers needed to drive >0.7m FR4 or 6m cable
– Increasing port bandwidth decreases port count
– EMI & signal integrity problematic

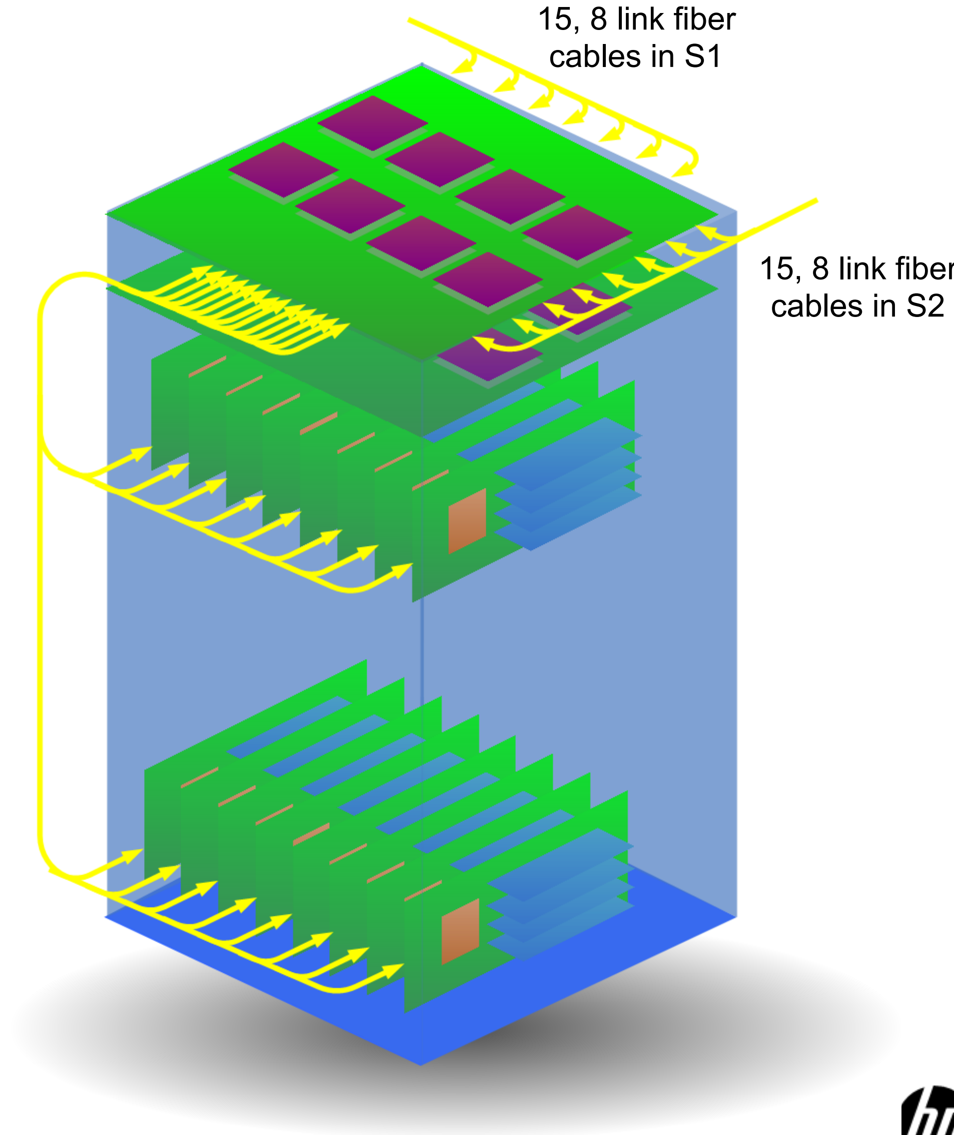
PHOTONIC

– Switch size unconstrained by device IO limits
– Port bandwidth scalable by increasing number of wavelengths
– Optical link ports can directly connect to anywhere within the data centre
– Greatly increased connector density, reduced cable bulk



TOPOLOGY TO PACKAGING MAPPING

- Building block is 128 node enclosure
- *L=3, S=[8,16,16], K=[2,1,1]*
- S_0 dimension within top of rack switch
- S_1 dimension along row of 16 rack
- S_2 dimension between 16 rows of racks
- 30 rack to rack connections per switch
- 16 nodes/switch/plane
- 64 way fiber cables (16 way with CWDM)



Memory

BLACKCOMB – DOE ADVANCED ARCH

ORNL, HP, Penn State, Michigan

– Nonvolatile memory devices, models, 3D stacking, experimental fab

– Architecture of NV memory

  - Resilience, wearout

– System architecture

  - Storage, checkpoints, DRAM replacement

– Programming

  - New abstractions for nonvolatile storage
  - New protocols for resilience in distributed memory code

– Applications

Memristor

Memristor path to NVRAM

– Compete commercially with Flash in ~ 3 years
– Solid-state drives soon thereafter
– Compete with DRAM in ~ 4-5 years
– On processor NV cache in ~ 4-5 years
– Compete with SRAM in ~5-6 years
– Universal NV memory and storage in 7-8 years

– Rethinking memory/storage hierarchy and interfaces now

Software

SOFTWARE TOPICS

- A resilient distributed programming model
  - programmer unaware
  - exploits local NVM for checkpoints
  - no cascading rollback
- System management software for
  - job control, guaranteed service levels
  - fault and bottleneck anticipation, discovery, diagnosis
  - cluster availability
  - networking
  - energy minimization
  - security
  - storage

Alternate processor

HP Labs ARM work

- What: ARM processors (and related SoCs) for servers
  - Microblades (ISCA 2008)
- Why:
  - Lower cost, better performance per watt
  - Comparable sustained application performance
- When: aggressively investigating possibilities

2018 EXAFLOP SYSTEM

PROCESSOR

- Silicon MCM compute complex
- ~10Tflops DP

SYSTEM

- ~33Gflops/W (system)
- ~100,000 nodes required
- 30MW (system)

MEMORY

- Direct attach 3D DRAM on silicon substrate.
- Photonically connected memory modules for expansion
- NV memory with low -power embedded compute elements
- Common architecture for persistent and volatile memory modules

INTERCONNECT

- On substrate CMOS nanophotonic communications fabric
- 640Gbps WDM photonic links
- HyperX topology
- Single mode fiber board to board and rack to rack

SOFTWARE THOUGHTS

– Big growth of commercial scale-out
  - Cloud data, email, etc
  - Sensor networks
  - Business data mining
  - Memory capacity, storage capacity, bandwidth, cost, power, interconnect bandwidth are all important in this space

– Scale
  - 10^8 threads
  - Memory capacity will be critical
  - Must remove scaling limits in OS, libraries, runtime

– Reliability
  - Solve the problem at the hardware, OS, runtime levels first.
  - Greater memory capacity is key (allow in-memory checkpoints)