Cloud Working Group

Co-Leads: Vijaykrishnan Narayanan and Klara Nahrstedt

Group Members: Sarita Adve, Sankar Basu, Luis Ceze, Tom Conte, Wilfried Haensch, Sharad Malik, Sayeef Salahuddin, Alan Seabaugh, Naresh Shanbhag, Lisa Theobald, Josep Torrellas, Cathy Yelick, Randy Bryant

Nanotechnology-inspired Information Processing Systems of the Future

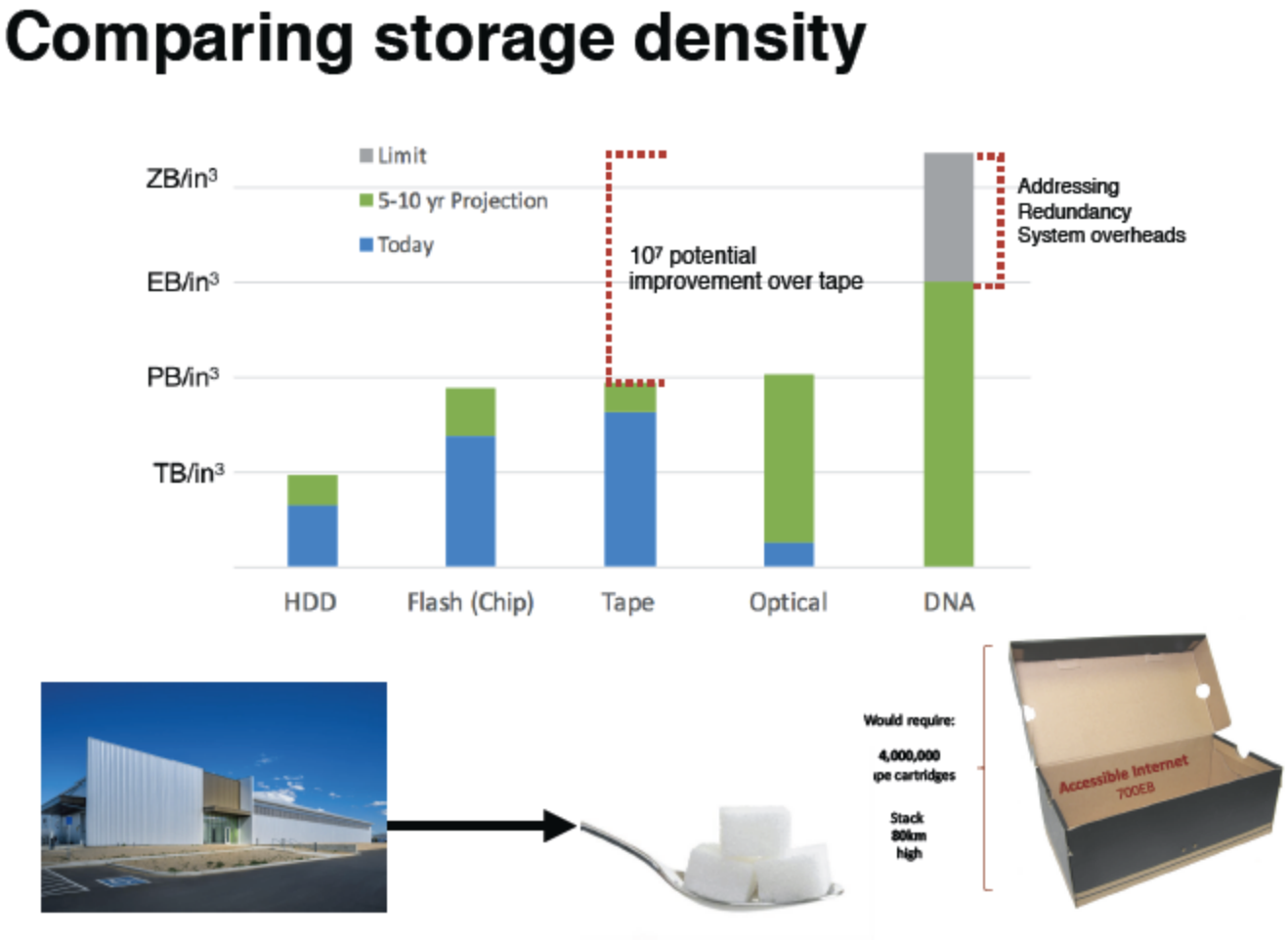
August 31-September 1, 2016

Fairmont Hotel, Washington DC.

Where is nanotechnology likely to play a key role ?

- Reduce size, cost, energy of data centers – DNA example
- New multifunctional devices – materials/devices innovation– memory and compute at same location – weight updates
- Reducing abstraction layers – map physical properties more directly to application computations – video analytics, graph algorithms, optimizations using oscillators
- Enable integration of memory and compute through reconfiguration (few ns) and fine grain integration (high density, metal –metal contact like connection)
- Centralized data centers make new technologies such as low temperature electronics attractive
- Rethink software stack that exploits new compute-memory integration enabled by nanotechnology to reduce overheads – virtual memory elimination
- Creating software to utilize unreliable or approximate fabrics to achieve overall system efficiency with flexible software-hardware boundaries – (Show potential for orders of magnitude improvement and software disruption challenges).
- Provide low latency (<10ns), high bandwidth (64 bytes/ns) non-volatile (10^15 cycles) memory
- Opportunities to enable new application domain centric data centers such as learning, inference, optimizations from specialized architectures.
- Energy-efficiency of edge devices will transform the needs on cloud and partitioning approaches

Some metrics



Cross Layer Research Opportunities and Challenges

Storage – Video – heterogeneous storage – bandwidth-driven – lot of data becomes “cold” soon – drives needs for new types of specialized memory in the design space.

File Systems, databases and query handling – enhanced by blurring between storage and memory – metadata and data layout overheads will reduce

Rethink software environment by new check pointing efficiency and unreliability of devices – Hadoop, Sparc

Networking switches have more in-node computations: edge, switches and the data center are a continuum.

Virtualization across different computational paradigms and substrates.

New programming models and hardware software interface to manage fine grain compute-memory integration under reliability constraints

New functions that can be done within memory and develop interfaces to memory that expose new functionality beyond read/write

Security (trust): Homomorphic encryption – compute on encrypted data; key management

Challenges

- Balancing functionality and ease of use with efficiencies of customization
- A flexible model that allows you to remove/add layers of abstraction to make tradeoffs between security, energy efficiency, etc
- Fewer layers of abstraction provides fewer protection layers and direct access to hardware – security hazard
- Legacy code – some edge nodes will have data formats and protocols that need to last even after transformations
- Ease of use and management is a key feature – how can the heterogeneity and new computational models continue to deliver his ease.
- Beware not to shift burden to software
- Software defined radios, Software define network motivate that hardware needs to be adaptive – reconfigurability – wireless links, nano-mechanical systems.
- **Demonstration systems are important – access to system prototypes refines the cross-layer discussions**

A Start

- Energy efficient computing
  - Can nanofunctions bridge the gap between application needs and device implementations?
  - Collective computing at scale
- Convergence of storage and memory
  - What memory design space does not have a solution -- a major quest
- Energy efficient communication
- Redefining interfaces –software-hardware interface
- Heterogeneity – a challenge for software and hardware

Beyond Moore and Von Neumann

- Cross Layer Innovation
  - CMOS scaling at the end of roadmap
  - Multiple walls – power, memory, I/O bandwidth
  - Specialization at various levels of design
    - Tap synergy between devices, circuits, architectures, computational models
  - Monolithic 3D integration – closer memory & compute and shorter & dense interconnect
  - Blur between memory and storage
  - What is the impact of new devices or computational models?
    - On compute, communication and memory
    - Algorithm has big impact on the data movement

Big Data

- Video data expected to dominate networks
  - 82% of Internet traffic by 2020 will be video (Cisco White Paper 2015-2020)
- Unstructured data types, irregular and streaming data accesses
  - Multi-Modal Data and Metadata including video/audio/text with location, timing, and other context information.
- Potential for inexact and approximate computing
  - Not all pixels are equal
- Workload consolidation and collaborative processing

Heterogeneity

- Specialization leads to diversity
- CPU, GPU, FPGAs + Neuromorphic processors, Quantum, Application Specific processors
- Software stack and programming models
- Impact of hardware diversity on operating system design and resource management
  - What additional OS layers are needed?

Sustainable Data Centers

- Software: End devices may have long life
  - Microscopes have lifetime of 10-15 years
  - Power-grid has end devices in place for 20 years
  - Smart City initiatives plan sensors installment for multiple years (too expensive to replace them because new HW/SW shows up).
- Software: Compatibility for long periods of time
  - Data format and computational models compatibility
  - Software lifecycles – OS upgrades
- Hardware
  - Scalable and adaptive specialization
  - Ability to integrate new technologies

- Energy, Area and Environmental Footprint for Data Centers is large
  - How do we design eco-friendly data centers?

- Sustainability of data, equipment, platforms is an issue.

Distributed Cloud + Edge intelligence

- Compressed Computing
  - Zettabyte Era – need to compress data and work on compressed streaming data
    - 2.3 ZB per year by 2020 (Cisco White Paper 2015-2020)
- Encrypted Computing
- Dynamic workload partitioning
  - Internet of Things will need edge computers to do partial computing before connecting to cloud
  - Cloudlets, edge computing, fog computing to assist clouds (cloud surrogates)
- Influence of new communication technologies on increased connectivity between edge devices
- Hierarchical cloud computing
  - Three tier cloud architecture – end device, edge device (cloudlet), cloud