Session C: Information technology impact on energy

Session Leader: Shekhar Borkar

Scribes: Tony Lentine & John Shalf

identify the key challenges to continuing Moore's Law from the systems perspective

- Earlier technical presentations covered this
  - Efficiency and performance of individual transistors (challenged by sub-threshold slope)
  - Data movement
  - Device density
- This is really about how we interpret moore's law
  - has been many things over time… density increases, energy efficiency improvements, cost improvements
- Why do we think things are worse now than they have been
  - Many more challenges to Moore's scaling than in past (increasing rapidly)
  - Diminishing return in extracting value per generation
- This is about improvement of value for each generation
  - Its NOT CMOS replacement (possibly CMOS augmentation)
  - Its about how do we increase (double) value per generation. (where value could be many different things)

Value Propositions

- Need a better definition for “Value” (see next slide)
  - Performance (HPC perspective)
  - Functionality (less well defined)
  - Capacity/storage
  - Efficiency (total IT energy cost)
    - Datacenters 2% and growing rapidly
    - All of IOT pushes closer to 10% (and fastest growing)
    - Other is societal cost of energy (even keeping it flat)

Metrics

- We really need quantitative metrics
  - value is good, but need something more specific and quantiative
  - Functionality (what consumers value) is ill defined, but still an important motivation

- Different metrics for different markets
  - IOT favors zero static losses (zero power at standby)
  - HPC favors performance and bandwidth density
  - Datacenter: bandwidth density and bandwidth distance
  - All have different fixed power envelopes

- Keren proposal (the unitary cube)
  - Unity is 1 byte, 1 byte/s, 1 op, for fixed unit area (volume), energy or cost
  - Can we increase value by 2x/year (or some compounded rate) with fixed energy or fixed cost (or both fixed)
  - Other value metric for N3XT-liked stacked memory is increased memory performance density in fixed area by stacking low-energy density memory (even without compute logic improvements)

what are opportunities for continuing technology scaling beyond 2025-2030

- We don’t know what the solution is in 10years, but
- The second question is really founded on “What does technology scaling MEAN after 2025”
  - What is scaled beyond 2025
  - Capacity per unit volume?
  - Performance of memory ?
  - Capacity per $?
  - Computation per $?

What should we do to organize ourselves to address these challenges.

For example, what would be the respective roles for industry and DOE in a public private partnership.

- There are other examples. We must first answer what are the gaps?
  - StarNet:
    - Covers individual topics
    - There are a couple of centers for architectures and one is materials
    - Does not include vertical integration

What does DOE have to offer?

- What DOE labs have at their disposal
  - DOE does integrated efforts
  - have expertise at all of these layers
  - Other agencies do focus areas, but not integration
- DOE can incorporate what comes from other agencies
  - DOE is an integrating hub (it is DOE’s mandate in NSCI)
- DOE can cover from basic science to manufacturing demonstration

Public/Private Partnerships

- Government should focus on precompetitive efforts where multiple companies can work together with Government
- Industry participants have said “industry alone will not solve these challenges”
  - Industry cannot justify high risk development projects to shareholders