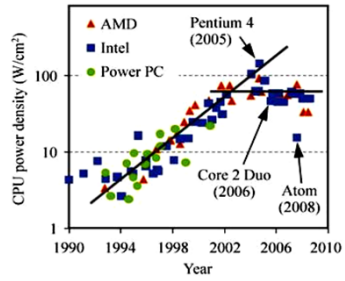
Elastic Fidelity: Trading-off Computational Accuracy for Energy Reduction

Georgios Tziantzioulis, Ali Murat Gok, S M Faisal, Ke Liu, Sourya Roy, Tyler Clemons, Nikos Hardavellas, Seda Ogrenci Memik, Srinivasan Parthasarathy

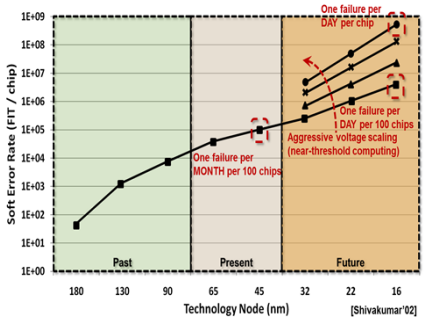
Trend: Increasing Power Dissipation



"Energy Dissipation and Transport in Nanoscale Devices", Eric Pop, Nano Research 3, 2010

- Power dissipation becoming an unmanageable problem
- Computing's energy consumption (2010): 408 TWh
- Google datacenters' energy needs: ¼ nuclear plant
- Datacenters carbon footprint: ~Czech Republic

Trend: Scaling Degrades Reliability

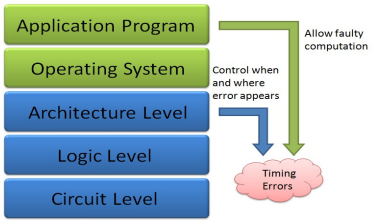


- Aggressive technology scaling degrades reliability
- Fabrication variation causes error surge
- Future computers will be built by inherently unreliable components

Thinking Beyond

Question: What if we relax the guarantees on reliability and intentionally allow components of the processor to fail sometimes with tolerable error rates?

- We can gain power savings at quadratic rate with voltage reduction.
- We have to accommodate these timing errors in the architecture and software layers.



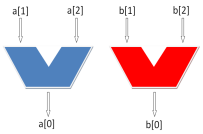
Elastic Fidelity Computing

- Different code and data segments of an application exhibit variable sensitivity to errors.
- Language constructs declare reliability requirements of each code/data segment.
- Hardware steers computation to components and adjusts their voltage to reach the reliability target.
- A program may still appear to execute correctly if it returns acceptable results from the user's perspective, even if there are inaccuracies in the computation.

Original compressed image

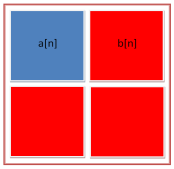
Decompressed image with 10% errors in lower 8 bits

Simple Example



Execution units (e.g., ALUs)

```
imprecise[25%] int a[N];
               int b[N];
. . .
a[0] = a[1] + a[2];
b[0] = b[1] + b[2];
. . .
```



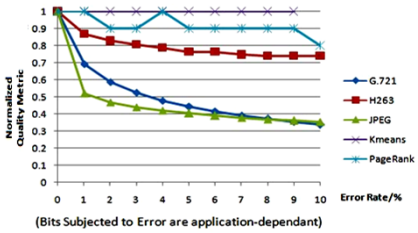
Data Storage (e.g., cache)

Voltage legend (color-coded)

- Fidelity requirement translated into voltage level
- Execute error-tolerant computation on low-voltage ALU
- Store error-tolerant data in low-voltage storage

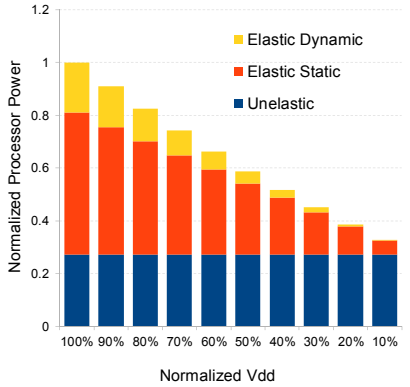
Methodology

- Software wrappers inject errors in computations at run time to simulate elastic-fidelity ALUs.
- Software wrappers inject errors in loaded data at run time to simulate elastic-fidelity storage.
- Calculate the output quality by comparing the results of error-free and error-injected runs.



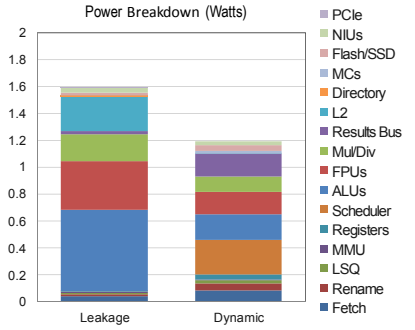
Output Quality with increasing error rates

Power Consumption (JPEG-D)



Assumed elastic components: ALUs, FPUs, Mul/Div, RF, L1D, L2

Power & Energy Savings (JPEG-D)



- Processor limit study @ 50% Vdd (Gem5/McPAT): 14% dynamic, 27% static power savings
- Cache limit study @ 70% Vdd (CACTI): 23% dynamic energy savings, 14% static power savings; even higher by foregoing unnecessary ECC

Summary

- Computing is unsustainable (energy, environment)
- Computing devices will sustain massive errors
- *Elastic Fidelity*: exploit the inherent error-tolerance of applications to lower the energy consumption, and withstand the massive errors of future computing.
  - Allow some data to be imprecise
  - Programming language constructs and ISA extensions pass the fidelity requirements from application to hardware
  - Hardware models adjust the voltage to maintain fidelity guarantees
  - The execution system steers imprecise computations to components with low voltage
  - Error-tolerant data are stored at low-voltage storage

Acceptable results, significant energy savings!

McCormick
Northwestern Engineering

THE OHIO STATE UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING