
Approximate Computing

Ravi Nair

IBM Thomas J. Watson Research Center

Yorktown Heights, NY

October 14, 2008

*Thanks to Dan Prener
for collaboration and useful discussions*

Growth of Data

- Volume of data produced increasing at an unprecedented rate
 - Sensors of various kinds
 - MEMS-based sensors/actuators increasing at 20% per annum
 - Results of scientific computations
 - NERSC experiencing doubling of data in storage every year
 - 3.5 Petabytes in 61 million files in 2007
 - Automation of human activities
 - Trading volume increasing at 19% per annum
 - Trading personnel increasing at 3%
 - Quote volume increasing at 55%
 - Breadth of use of computation
 - Pervasive use of mobile devices
 - Increased computational capability of these devices

Streaming Systems

- Complex Event Processing (CEP)
 - Goal: Identifying complex relationships between events in an event cloud
- Event Stream Processing (ESP)
 - CEP where the event cloud is represented in streams of event data
 - Applications
 - Algorithmic trading in financial services
 - RFID event processing applications
 - Fraud detection
 - Process monitoring
 - Location-based services in telecommunications
- Analysis of data on the fly
- Fundamentally different from traditional database systems
 - Static data, dynamic query vs.
 - Static query, dynamic data
- System has to be designed to tolerate variations in the incoming data rate

Energy Efficiency: The Driver for Approximate Computing

- In many applications, approximate results can be tolerated
 - Not of course calculations involving your bank account
- Special-purpose solutions giving approximate but tolerable results tend to be energy-efficient
 - Several inefficiencies in today's "exact" computing paradigm
- Commercial applications such as data-mining, search, and analytics consuming increasing fraction of cycles
 - These applications can tolerate some imperfection in the results
- Same for media applications – photo, video, audio
- Applications that have ephemeral output tolerate approximation
- The common man expects computers to provide answer similar to those of an expert human being
 - Give me a route that, in your best opinion, gets me quickly to Manhattan
 - Not, give me the fastest route to Manhattan

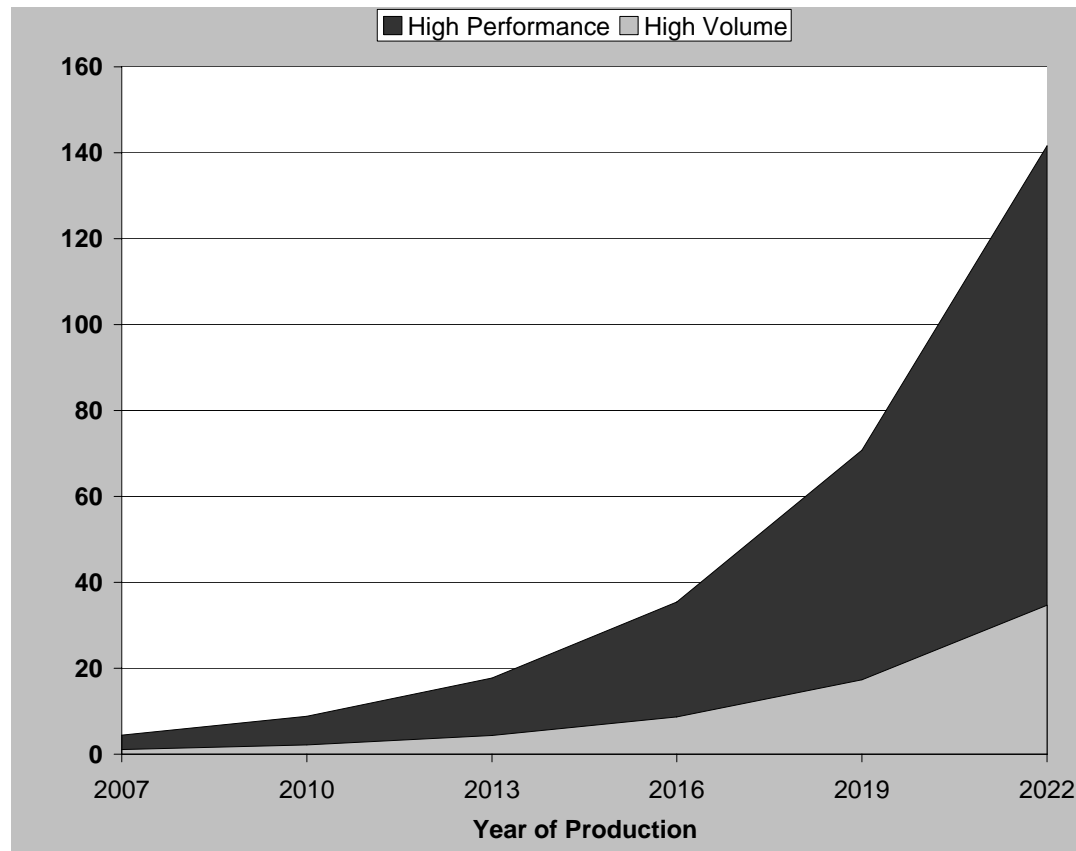
Approximate Techniques already in use

- Many of Google's Search techniques can be considered approximate
 - Does not work with a coherent, up-to-date database of Web pages
 - Query sent to many computers, some of which may fail during computation
 - Lack of "prompt" response from some computer causes query to sent to one or more other computers
- Google's Map-Reduce programming paradigm has provision for ignoring consistently failing records
 - Dropping an occasional record does not affect computations that are statistical in nature
- The deployment of tens of thousands of commodity processors over inexpensive networks requires a fundamental rethinking of the algorithm

IBM Roadrunner for Los Alamos National Lab

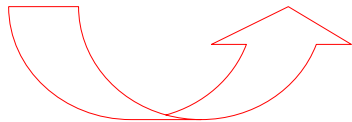
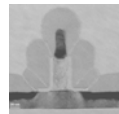
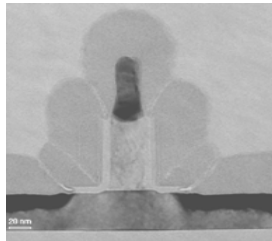
- 1 Petaflops
- 6562 Dual-core AMD Opteron chips
- 12240 Cell chips (used in Sony Playstation 3)
- 98 Terabytes of memory
- 278 refrigerator-sized racks
- 2.35 MW of power

ITRS 2007 Roadmap – Transistors per Chip



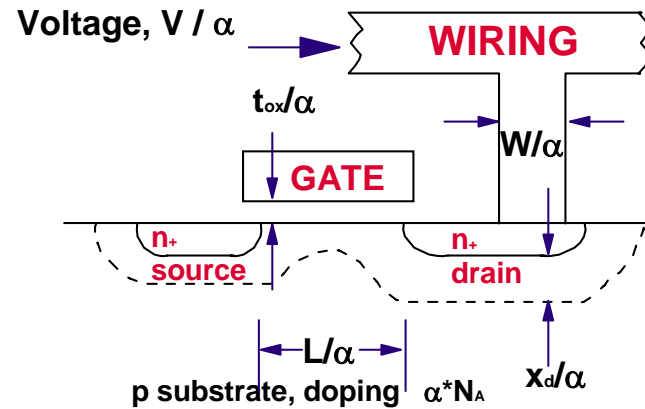
CMOS Scaling: Dennard's Theory

Scaled technology generations



Smaller
Faster
Lower
Power

Scaled Device



Dennard, 84

SCALING:

Voltage: V/α
 Oxide: t_{ox} / α
 Wire width: W/α
 Gate width: L/α
 Diffusion: x_d / α
 Substrate: $\alpha * N_A$

RESULTS:

Higher Density: $\sim \alpha^2$
 Higher Speed: $\sim \alpha$
 Power/ckt: $\sim 1/\alpha^2$
 Power Density: $\sim \text{Constant}$

Chart - courtesy G. Shahidi

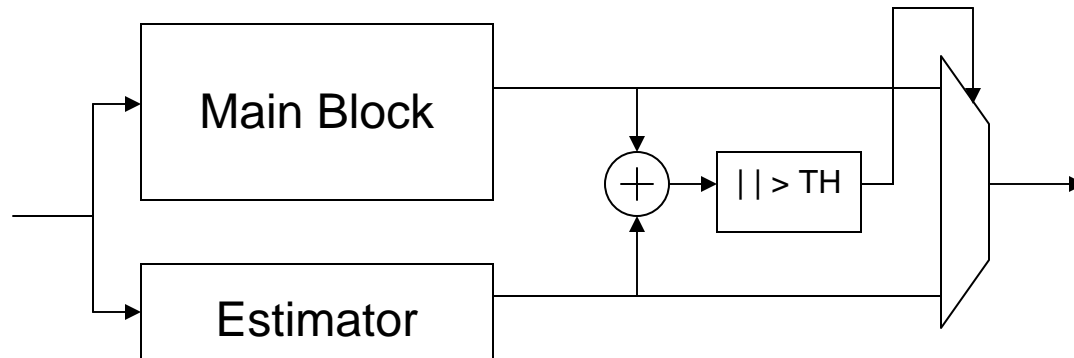
Performance Unpredictability

- Small geometry leads to
 - Process variability
 - Hence performance variability
 - Greater unreliability
 - Need to tolerate failures

Three Levels of Techniques for Approximate Computing

- Algorithmic Level
 - Tolerate errors in hardware
 - Take algorithms up above the brute-force particle level
 -
- Architecture Level
 - Architecture = Software/Hardware Interface
 - Specify precision tolerance
 - Specify incoherence tolerance
 -
- Implementation Level
 - Incorporating redundancy to combat unreliability is likely to be inefficient for most new technologies
 - Instead, implement larger functions that compromise the exactness of the solution

Naresh Shanbag's Algorithmic Noise Tolerance



- Employs statistical signal processing techniques
- Main Block is designed for average case
 - Makes intermittent errors
- Estimator approximates Main Block output
- Error-correction: Compare and replace
- Approximate, because error is not accurately known

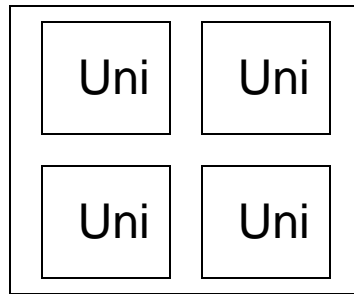
Utilization

- Maximal utilization of transistors not as important
- Scarce resources are
 - Power
 - Bandwidth
- Use real estate for
 - Different forms of accelerators
 - Different types of cores
- Turn on only those accelerators or cores that are needed
 - Within chip power budget

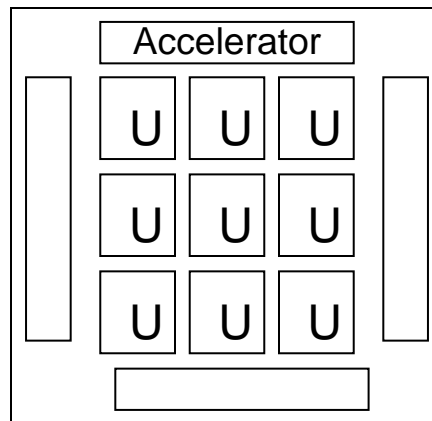
Possible Evolution of Computer Chips



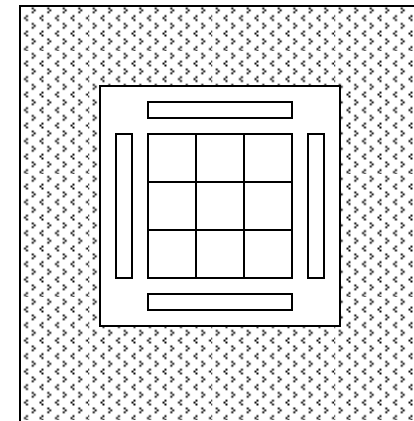
Single-core chip



Multiprocessor chip

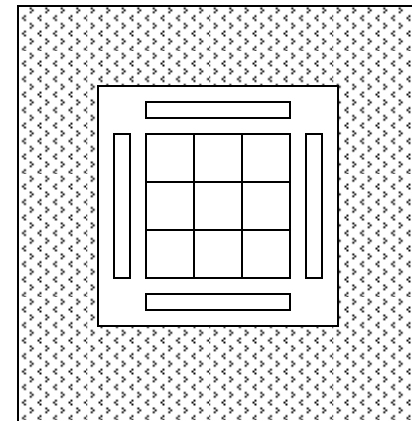
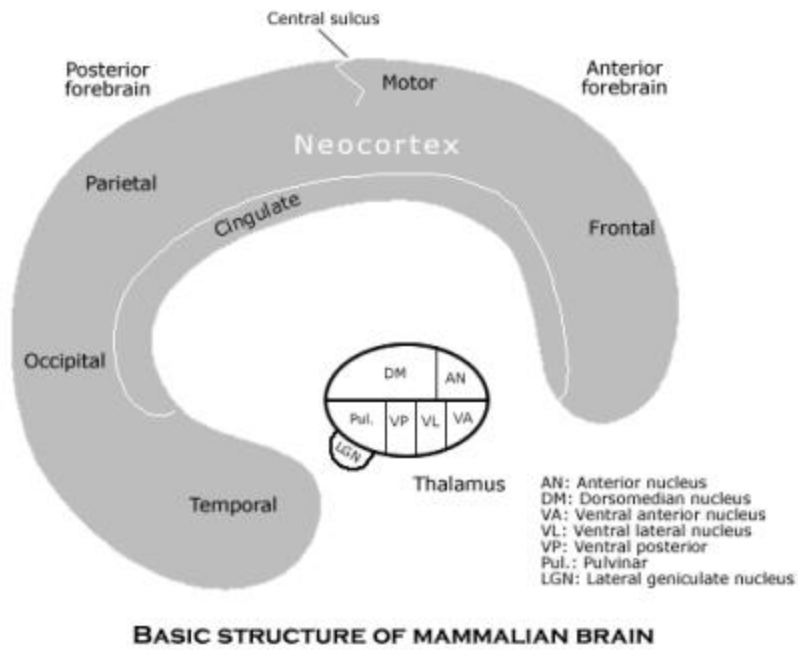


With energy-efficient accelerators



"Intelligence" through learning superstructure

The Brain



From CRA Grand Challenges 2005 Whitepaper

It is probable that new technologies --nanotechnologies (e.g., carbon nanotubes [CNTs], quantum cellular automata [QCAs], etc.) -- will demand new models of computation

.....

These new technologies will certainly require innovations across the entire computational stack, including microarchitectures, execution models, instruction sets, compilation algorithms, languages, and programming models

.....

The most powerful computing machine of all is the human brain. Is it possible to design and implement an architecture that mimics the way the human brain works?

Conclusions

- Data explosion taking place
- Types of computation done on data undergoing qualitative change

- Density scaling continuing at historic rates
- Power and reliability becoming serious problems

- Approximate computing lies right at the intersection of both these trends
- Techniques to support approximate computing techniques will have to span all levels of the computing paradigm

- Continue to gain understanding of the mechanisms of the human brain – a remarkably efficient approximate computing appliance