Quantum Tunneling Through The Memory Wall

The Case for Computing in Flash in the Age of Big Data

Kevin Gomez – Seagate, SSD System Architecture and Advanced Development
kevin.gomez@seagate.com
Peng Li – University of Minnesota, Center for Research in Intelligent Storage
lipeng@umn.edu

Collaborators

Robert Thibadeau, Chief Scientist, Wave Systems

Dave Touretzsky, Center for the Neural Basis of Cognition at CMU

Tom Mitchell, Chair of Machine Learning Department at CMU

Terence Sejnowski, Head, Computational Neurobiology Laboratory, Salk Institute

Dan Hammerstrom, Program Manager Microsystems Technology Office (MTO), DARPA

David Lilja, Head ECE University of Minnesota

Outline

- Memory Wall
- Future Computing
- Why NAND Flash
- Emerging Big Data applications
- Simulation Results
- Summary

Memory Wall

Hitting the Memory Wall: Implications of the Obvious

Wm. A. Wulf
Sally A. McKee

Department of Computer Science
University of Virginia
{wulf | mckee}@virginia.edu

December 1994

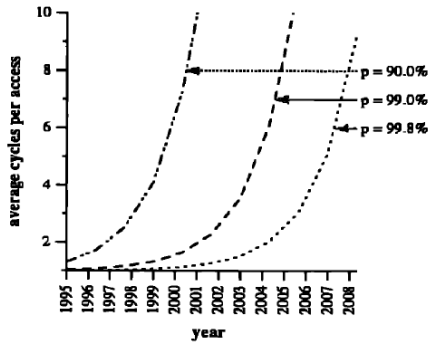
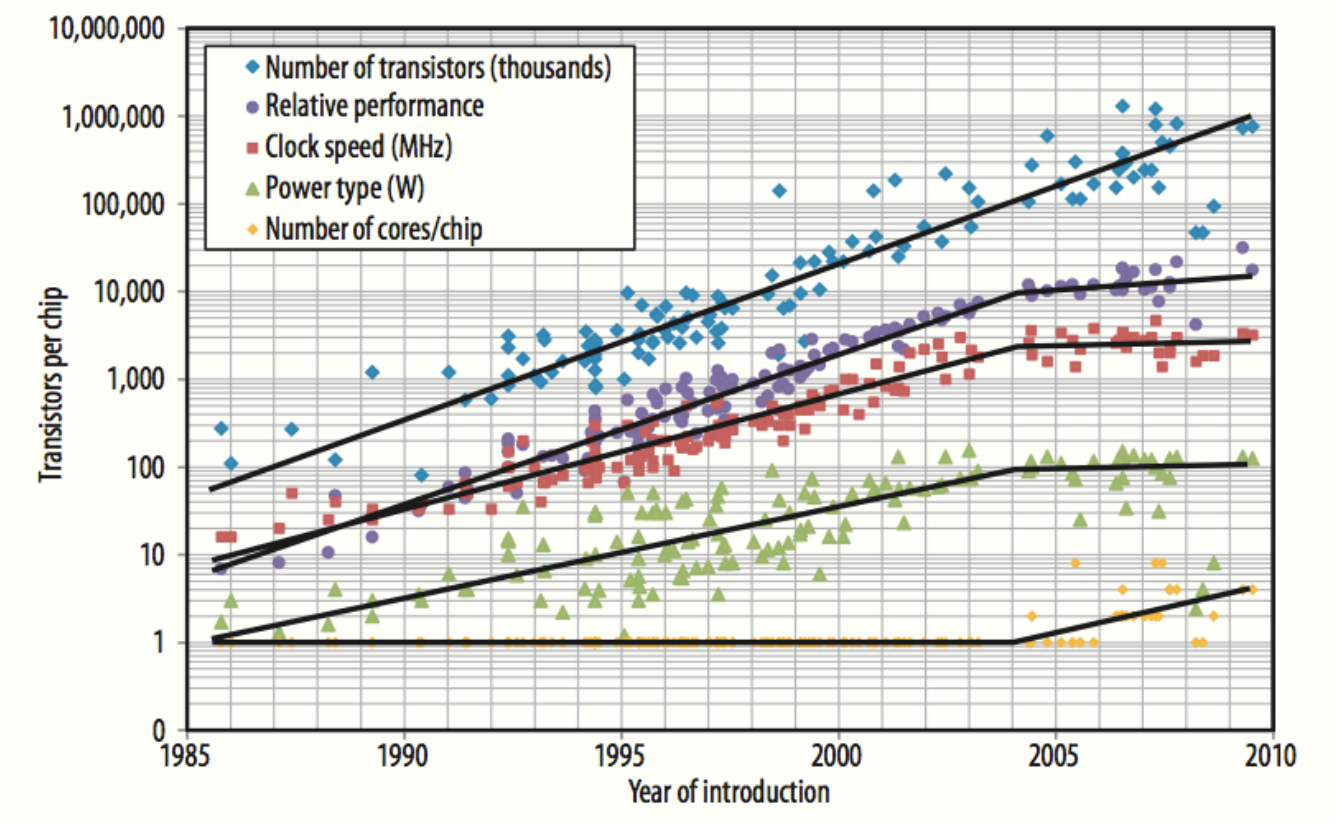


Figure 3 Average Access Cost for 80% Annual Increase in Processor Performance

In 1994 Wulf and McKee pointed out the implications of processor and memory performance progressing exponentially but with differing rates (~50%/yr for processors vs 7%/yr for memory) – causing an exponentially increasing gap which would lead to the end of single thread processor performance progress by 2008

Predictions were largely accurate



Transistors, frequency, power, performance, and processor cores over time. The original Moore's law projection of increasing transistors per chip remains unabated even as performance has stalled.

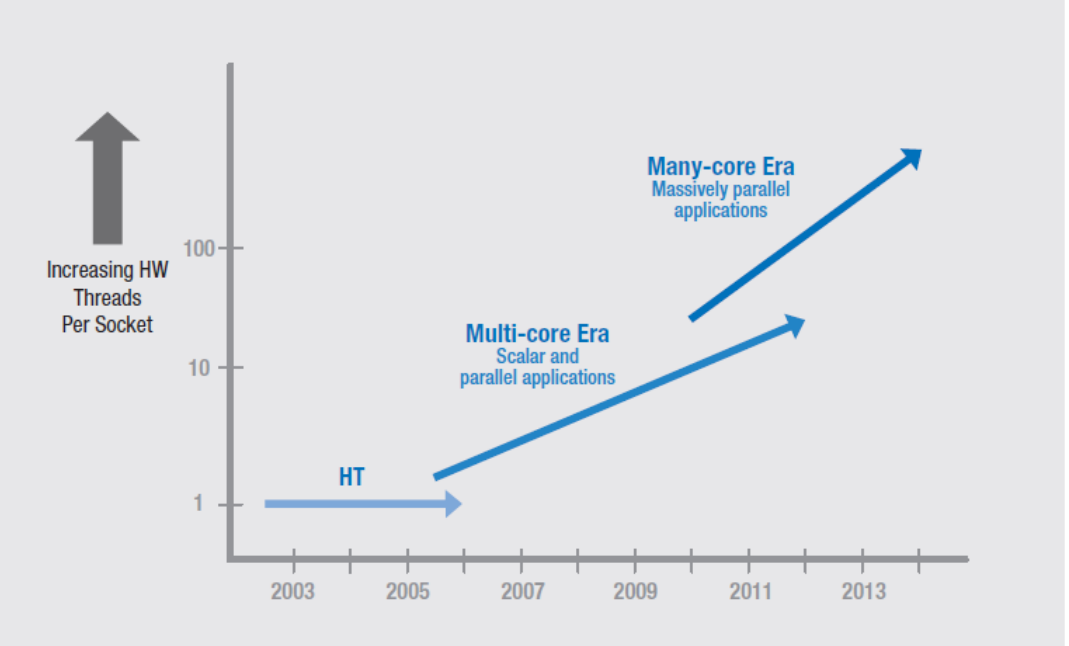
"Computing Performance: Game Over or Next Level?"
Fuller et al, 2011

Power Wall

The Power Wall was Hit around 2004 due to a breakdown in Dennard Scaling
Where Power had previously scaled as $1/L^3$
at ~ 2004 was limited to scaling as 1/L

Vdd scaling -> lower Vth -> exponential increase in leakage
This caused diminishing returns on pushing single-thread processor architectures

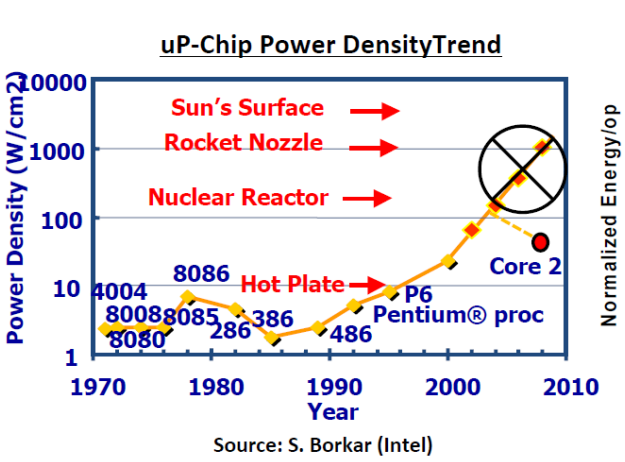
Intel Reaction in 2005 – “Platform 2015 – White Paper – Intel Processor and Platform Evolution for the Next Decade”
use Moore scaling to double cores every generation

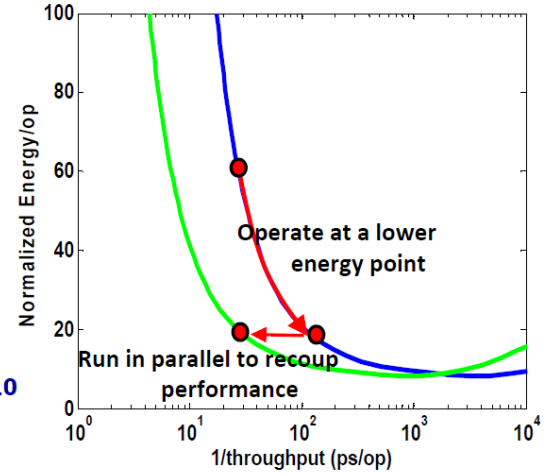


Increased Parallelism has forced the emergence of New Computing paradigms

Due to Amdahl's Law - simply parallelizing legacy applications has rarely yielded proportionate performance increases

Parallelism





- **Parallelism is the main technique to improve system performance under a power budget.**

S. Borkar, Intel

Computing Performance

Game Over or Next Level

The era of sequential computing must give way to a new era in which parallelism holds the forefront. There is no guarantee we can make parallel computing as common and easy to use as yesterday's sequential single-processor computer systems, but **unless we aggressively pursue the efforts** suggested by the CSTB* committee's recommendations, **it will be *game over* for growth in computing performance**.

- *Samuel H. Fuller "Computing Performance – Game Over or Next Level" 2011 and Chair - Committee on Sustaining Growth in Computing Performance (CSTB)**

IEEE Computer Society – March 2013

"It's time to rethink the entire approach to computation..

…We have been using the same models for computation since the inception of computing. We've tweaked and optimized every level of the stack, but to meet today's challenges, everything has to be on the table.
This will require a serious, cross-discipline conversation among domain experts."

Tom Conte, vice president of the IEEE Computer Society

IEEE - Rebooting Computing Working Group

Formed 2013

"Revamping computing is not something that any organization or company can undertake by itself. IEEE has societies and councils engaged in almost every aspect of computing, so our organization is the natural place to take on these tasks.. **The goal is to completely rethink computing, from devices to circuits to architecture and software** …IEEE will be the catalyst to spawn new thinking."

Elie Track, chair "IEEE Rebooting Computing Group"

Dark Silicon

However increasing parallelism simply delays the problem

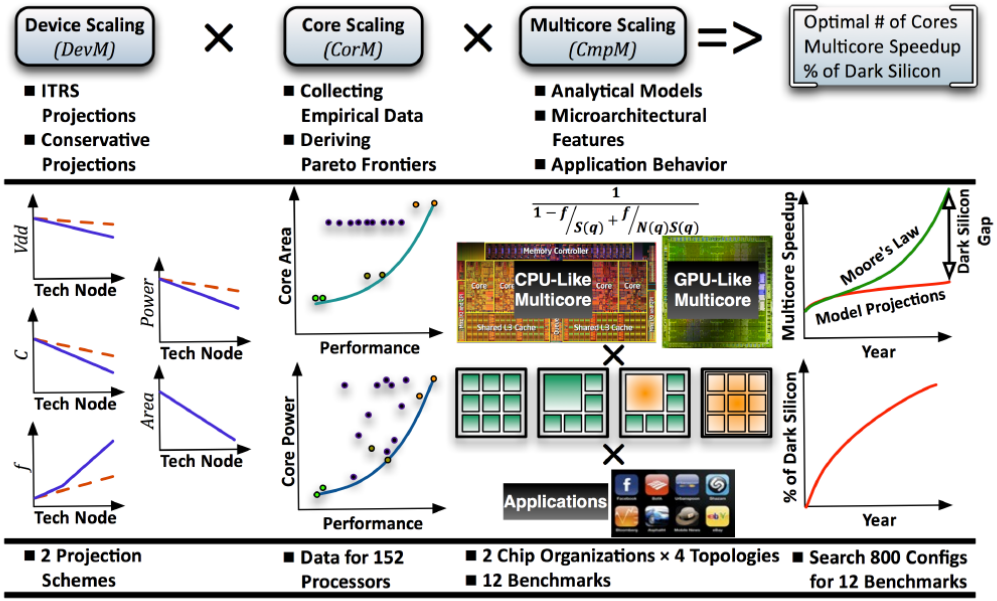


Figure 1: Overview of the models and the methodology

"..[Modeling] must consider devices, core microarchitectures, chip organization, and benchmark characteristics, applying area and power limits at each technology node. This paper considers all those factors together, projecting upper-bound performance achievable through multicore scaling, and measuring the effects of non-ideal device scaling, including the percentage of "dark silicon" (transistor under-utilization) on future multicore chips. Additional projections include best core organization, best chip-level topology, and optimal number of cores.

"Dark silicon and the end of multicore scaling"
H Esmaeilzadeh, et al … (ISCA), 2011

Dark Silicon

Conclusions

"….Dennard scaling's failure led the industry to race down the multicore path, which for some time permitted performance scaling for parallel and multitasked workloads, permitting the economics of process scaling to hold..

..An essential question is how much more performance can be extracted from the multicore path in the near future.

…This paper combined technology scaling models, performance models, and empirical results from parallel workloads to answer that question and estimate the remaining performance available from multicore scaling. Using PARSEC benchmarks and ITRS scaling projections, this study predicts best-case average speedup of 7.9 times between now and 2024 at 8 nm. That result translates into a 16% annual performance gain, for highly parallel workloads…"

Heterogeneous Architectures

Lots of efficient H/W automation – powered off most of the time

As Moore continues to increase the number of transistors on silicon at a scale of $1/L^2$ while power is only decreasing as 1/L …

… we can afford to 'overprovision' the chip – i.e. use the TDP (total die power budget) using just a subset of the chip's resources – for example use the entire budget on compute while shutting down global on-chip communication resources.

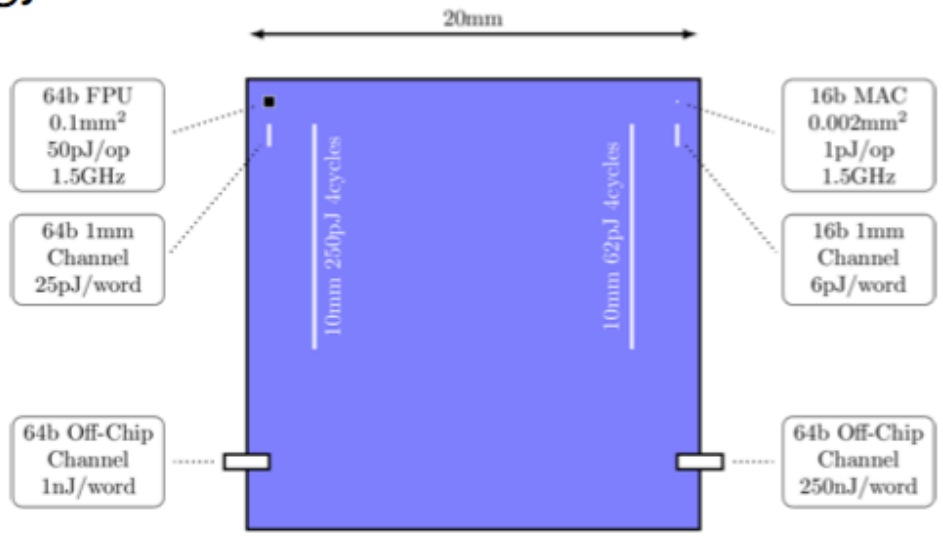
Enables peak performance (using all available power) on diverse workloads.

This may signal that the right time for Reconfigurable Computing has arrived – specialized hardware acceleration, powered off most of the time.

The Cost of Moving Data

- Key limitation in GPU performance → power consumption
- Ops vs. data transfer over large distances on/off chip
- Compare area and energy:
  - 16-bit MAC
  - 64-bit FPU
  - channels

- FP: 10x more energy-efficient than moving a word .5 die length (e.g. from the LL-cache)
- 16b MAC: 100x
- Off-chip: 40x more!
- ALSO: Wire delays do not scale as fast as transistor speeds



Bill Dally, International Conference on Supercomputing 2010

"Architecture at the end of Moore", Stefanos Kaxiras,, 2012

Energy cost now is dominated by data movement

Table 1. Technology and circuit projections for processor chip components.

	2010	2017	
Process technology	**40 nm**	**10 nm, high frequency**	**10 nm, low voltage**
V_{DD} (nominal)	0.9 V	0.75 V	0.65 V
Frequency target	1.6 GHz	2.5 GHz	2 GHz
Double-precision fused-multiply add (DFMA) energy	50 picojoules (pJ)	8.7 pJ	6.5 pJ
64-bit read from an 8-Kbyte static RAM (SRAM)	14 pJ	2.4 pJ	1.8 pJ
Wire energy (per transition)	240 femtojoules (fJ) per bit per mm	150 fJ/bit/mm	115 fJ/bit/mm
Wire energy (256 bits, 10 mm)	310 pJ	200 pJ	150 pJ

Table 2. Projected bandwidth and energy for main-memory DRAM.

	2010	2017
DRAM process technology	**45 nm**	**16 nm**
DRAM interface pin bandwidth	4 Gbps	50 Gbps
DRAM interface energy	20 to 30 pJ/bit	2 pJ/bit
DRAM access energy[6]	8 to 15 pJ/bit	2.5 pJ/bit

Locality = Efficiency

Scaling trends in Tables 1 and 2 suggest

Energy required to;

Read three 64-bit source and write one destination operand is equivalent to -

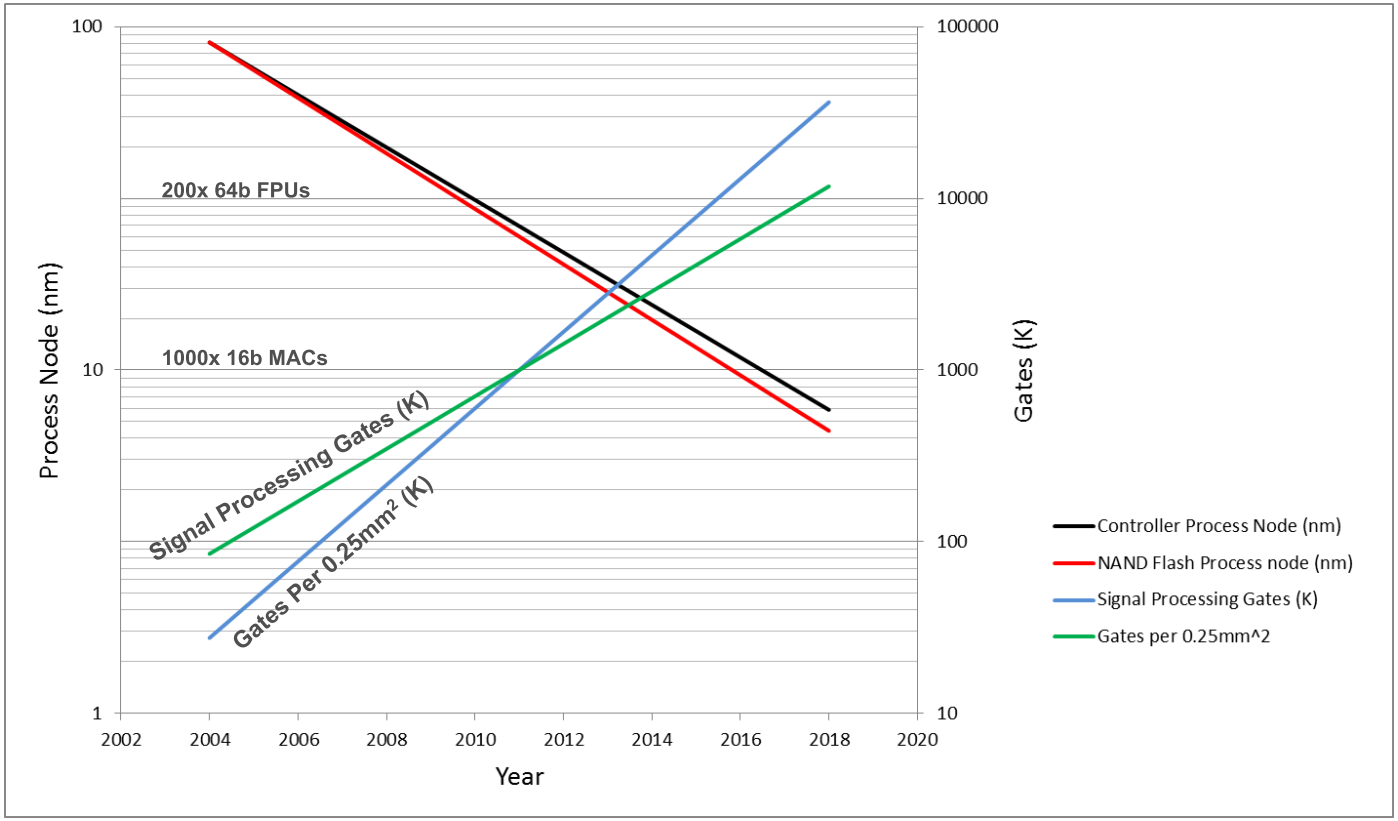
One double-precision floating point multiply accumulate @ ~56pJ

Accessing operands from 10mm away costs 6X more @ ~300pJ

Accessing from external DRAM is 200X at @ 10nJ

"GPUS AND THE FUTURE OF PARALLEL COMPUTING", William Dally IEEE Micro 2011

NAND process node and corresponding channel complexity needed to maintain system error rate



If the NAND controller is pad limited it would be feasible to increase die utilization and add compute functionality with zero increase in die size

Why NAND and not other NVM technologies

- NAND is a block device and requires a significant and growing investment in signal processing to enable it's continued scaling
- This signal processing overhead is best situated close to NAND to minimize the energy cost of data movement
- NAND has no delusions of being a DRAM replacement like PCM or STTRAM with low-latency and close to byte addressable architectures which will not tolerate any significant signal processing overhead
- **It is not about the technology – it's the economics -** SSDs exist due to the much larger demand for consumer grade NAND devices for the smartphone, tablet, SD Card and USB memory markets.
- Storage-Compute likewise will succeed or fail purely on economics ($/op, J/op) not technology

ITRS – Technology Trends

for DRAM and FLASH Memory

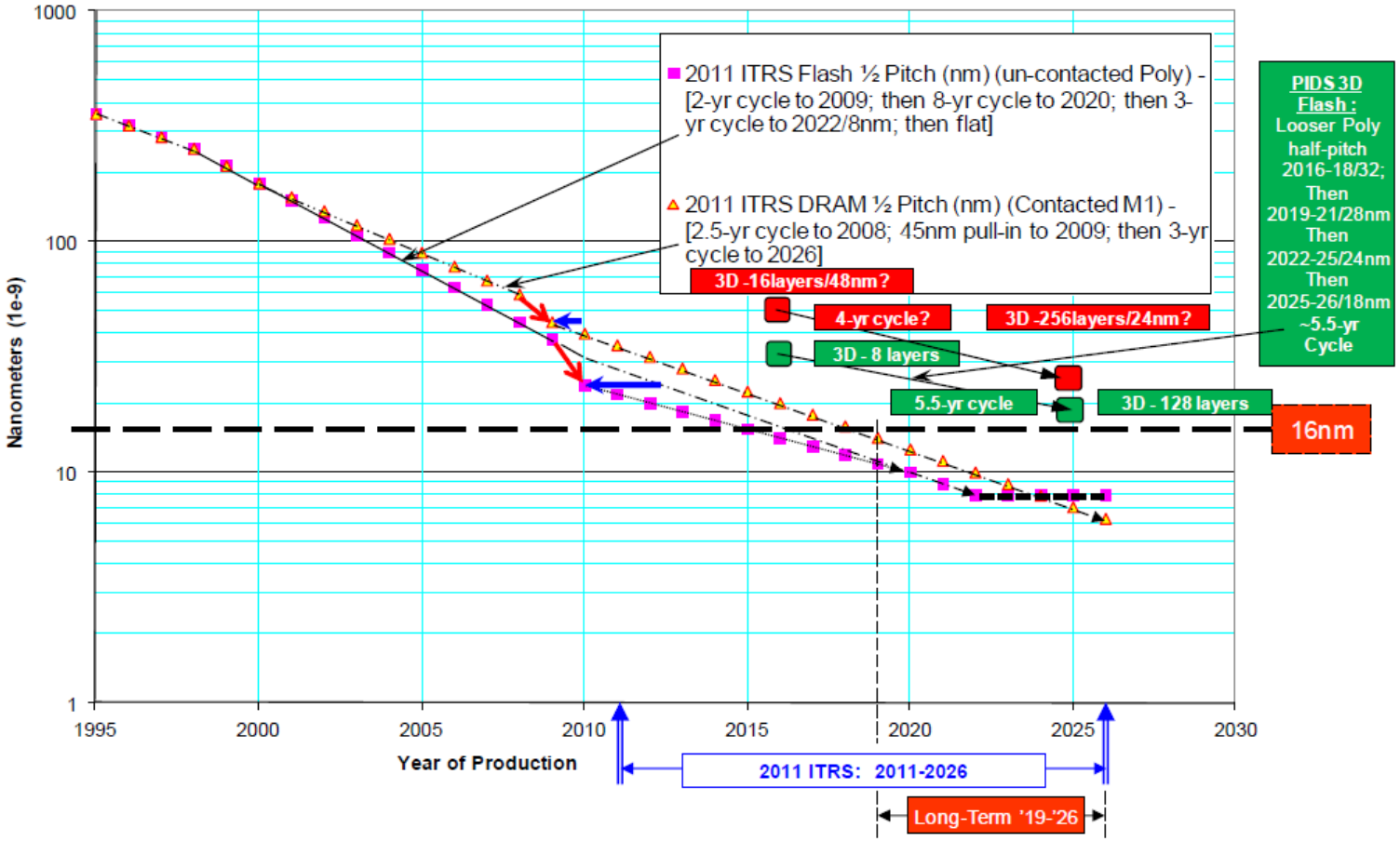
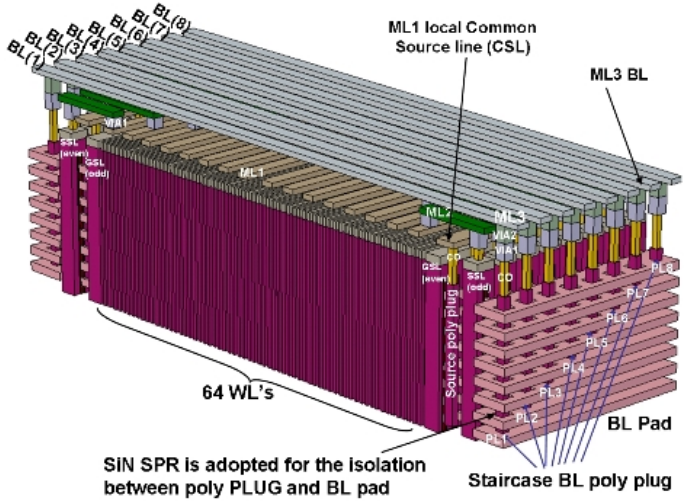


Figure 10 2011 ITRS—DRAM and Flash Memory Half Pitch Trends

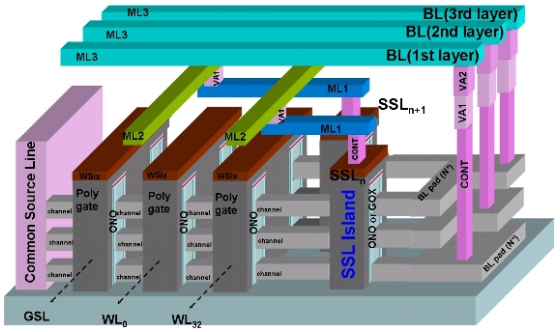
NAND will continue to scale… vertically

… and dominate the mind-share of the best memory technologists

NAND Flash														
Year of Production	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*	*2021*	*2022*	*2023*	*2024*	*2025*
Uncontacted poly 1/2 pitch (nm)	20	18	17	15	14	13	12	11	10	9	8	8	8	8
Number of word lines in one NAND string	64	64	64	64	64	64	64	64	64	64	64	64	64	64
Dominant Cell type	FG	FG	FG/CT	FG/CT	CT-3D	CT-3D	CT-3D	CT-3D	CT-3D	CT-3D	CT-3D	CT-3D	CT-3D	CT-3D
Maximum number of bits per chip (SLC/MLC)					128G / 256G	256G / 512G	256G / 512G	512G / 1T	512G / 1T	512G / 1T	1T / 2T	1T / 2T	1T / 2T	2T / 4T
Minimum array 1/2 pitch - F(nm) [15]					32nm	32nm	32nm	28nm	28nm	28nm	24nm	24nm	24nm	18nm
Number of 3D layers for array at minimum 1/2 array pitch [16]					8	16	32	32	64	64	98	98	98	128



- ITRS Winter Public Conference Dec 2012 Hsinchu, Taiwan



ITRS – Winter Conference, Dec 2012

"Take-away Comments"

Logic:

- No theoretical scaling limit seen yet for Si (to 2026, gate length ~6 nm).
- Power is the limiting factor, not speed. Device speed requirement is relaxed from circuit perspectives.
- Alternative channel III-V/Ge can offer lower power with similar speed.
- Low V_{dd} near end of roadmap (~0.5 V) posts noise/variability challenges.
- Series resistance can be a practical limitation.

DRAM:

- Capacitor scaling increasingly difficult.
- $4F^2$ is the limit for cell size.

NVM:

- Many cell versions:
  - 3-terminal (charge-based): Floating-gate and charge-trapping FETs still dominate. 3-D projected.
  - 2-terminal (non-charge-based): FeRAM, PCRAM, MRAM, STT-RAM, for more diverse applications. Efficient selection device needs to be developed and integrated.

End

Managed NAND

Moving Flash endurance optimization processing and channels into Flash package

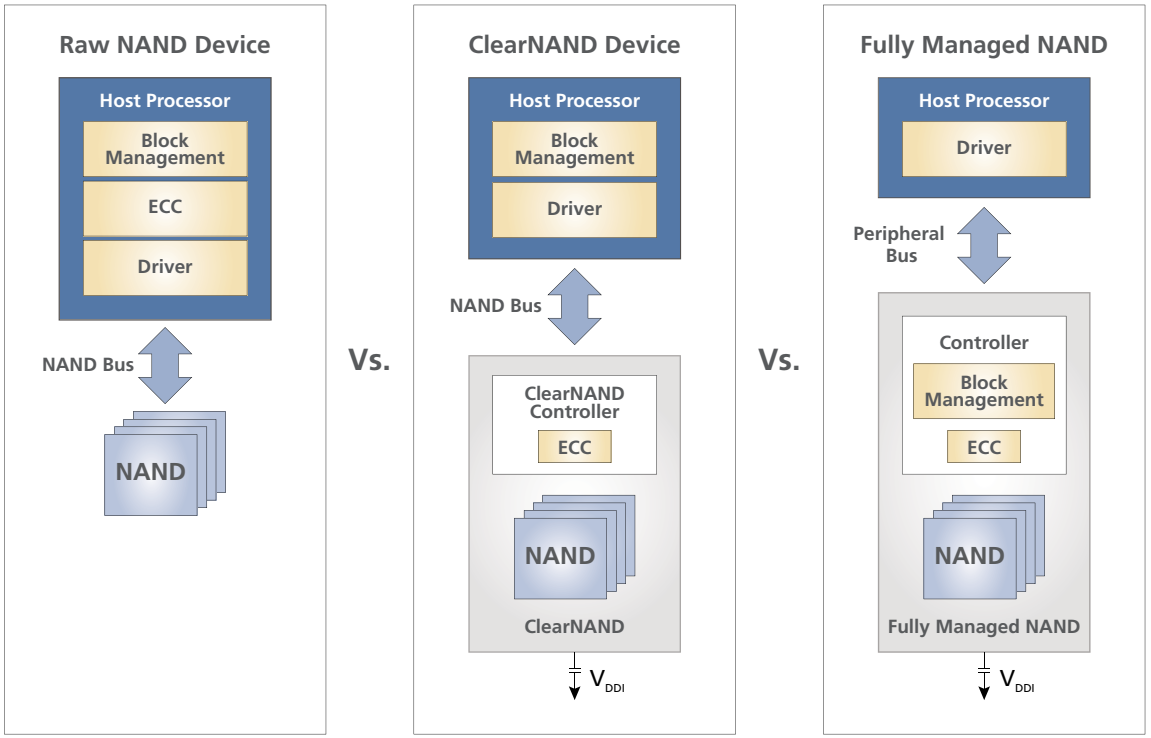
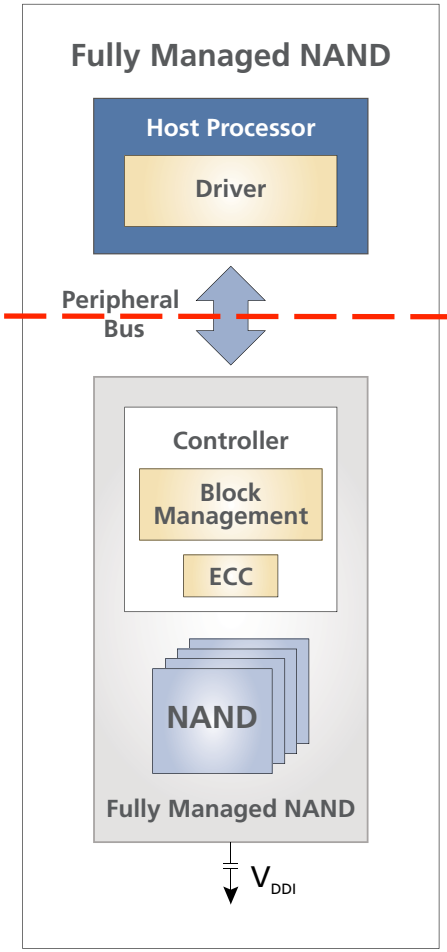


Image – ClearNAND Flyer - Micron

Increasing complexity of ECC/ channel processing integrated directly into NAND Flash packages

Managed NAND

A construct for IP isolation



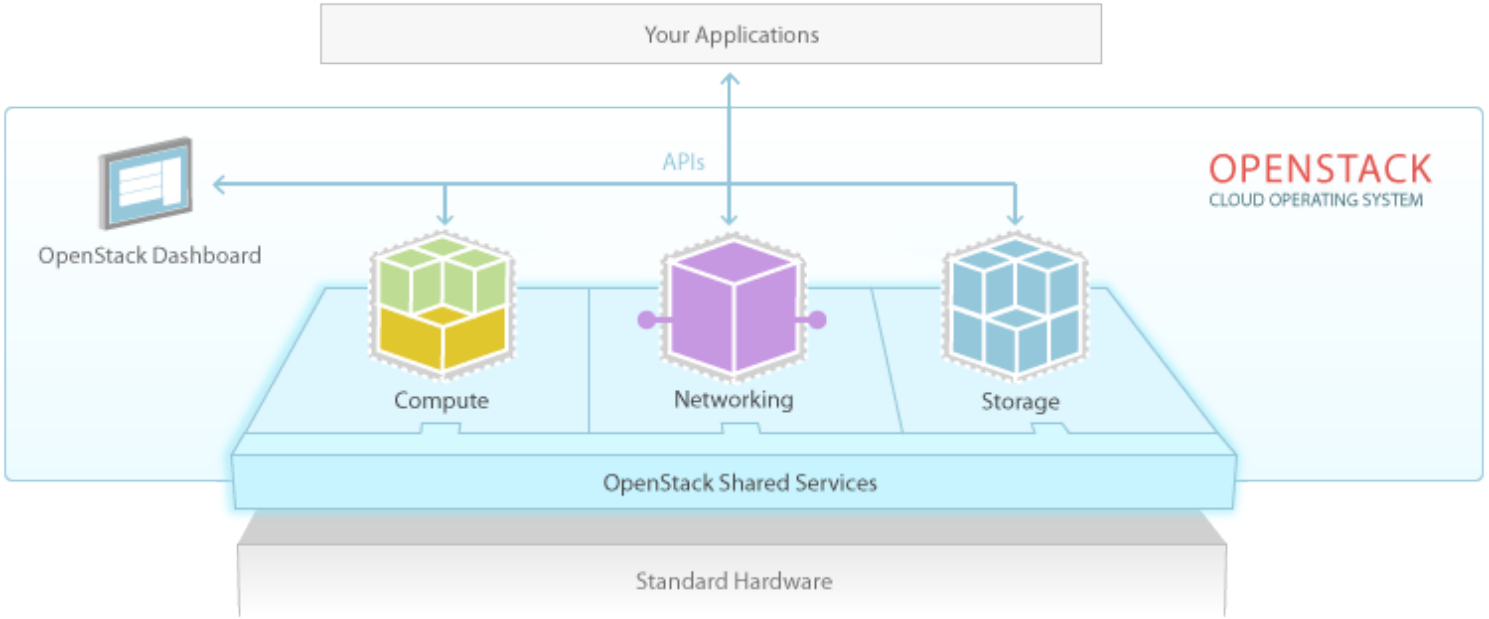
FUSION-IO
Texas Memory Systems
intel
SanDisk
WhipTail
VIRIDENT
NIMBUS DATA
Seagate
Violin MEMORY
BITMICRO
SMART STORAGE SYSTEMS
sTec
OCZ Technology
PURESTORAGE
EMC²
LSI Storage. Networking. Accelerated.
Micron
SKYERA SOLID STATE STORAGE SYSTEMS

Intellectual Property Firewall

SanDisk TOSHIBA
linkmedia Link_A_Media Devices
SK hynix
intel
Micron
SAMSUNG
SAMSUNG

The Virtualization Revolution

OpenStack - Open source software for building private and public clouds.



http://www.openstack.org

OpenStack/Open Compute – the launch vehicle for application development on SSDs used as Storage Compute Elements

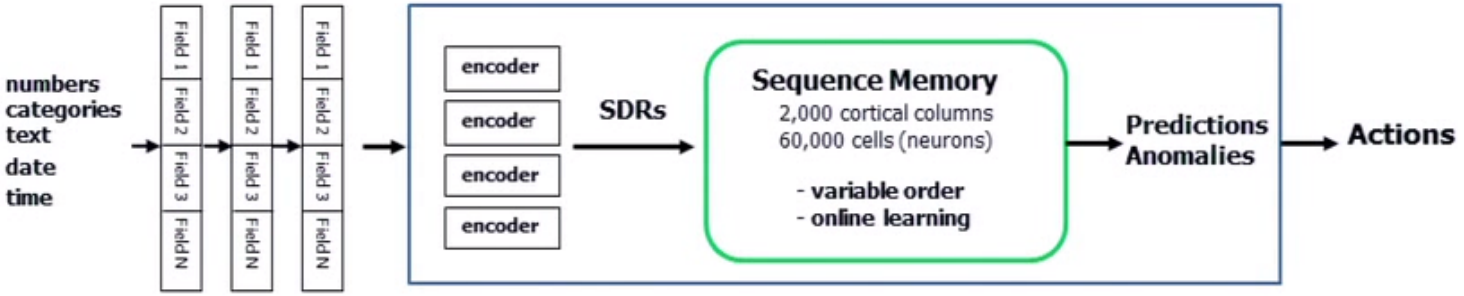
Big Data – application areas

Some Application Areas

- content based image retrieval
- semantic tagging of streaming multimedia
- data intensive scientific computing
- on-line machine learning from streaming data
- cloud compute optimized for Numenta's Grok Engine
- on-line spatiotemporal clustering and classification of trading data
- hierarchical-temporal memory model based algorithmic trading in financial markets

Big Data Application Example – Numenta's Grok

An Engine for Acting on Data Streams



User

Define problem

Stream data

Grok

Creates models

Learns

- spatial/temporal patterns

Outputs

- predictions
  anomalies

Applications

Energy pricing

Energy demand

Product forecasting

Anomaly detection

Server loads

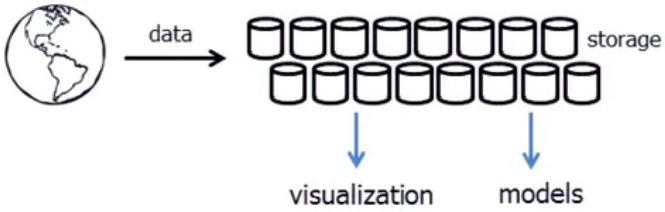
"Building brains to understand the world's data"
Jeff Hawkins, Feb 2013

Big Data Application Example – Numenta's Grok

An Engine for Acting on Data Streams

Data

Today



Challenges

People, not automated
Model obsolescence
Streaming data

Tomorrow

Key criteria

Automated model creation (B's of models)
Continuous learning
Temporal and spatial patterns

"..No Storage Required.."

"Building brains to understand the world's data"
Jeff Hawkins, Feb 2013

Data Intensive Scientific Computation

Trends in computation, communication and storage and the consequences for data-intensive science

Simone Ferlin Oliveira, ferlin@nm.ifi.lmu.de
Karl Fürlinger, fuerling@nm.ifi.lmu.de
Dieter Kranzlmüller, kranzlmueller@nm.ifi.lmu.de
Ludwig-Maximilians-Universität
Munich Network Management (MNM) Team
Oettingenstraße 67, 80538, Munich, Germany

***Keywords*-data-intensive science; big data; trends**

***Abstract*—The way we are doing science is changing: Data analysis and computation modeling became tightly coupled. Divergent technological trends for computer processors, storage and memory and communication systems showed to be a real challenge in performance of current computing systems. In this paper we analyze the trends that influence computer performance, point out the technical challenges and introduce our vision in developing a guideline to an optimum distribution of computer resources addressing primarily data transmission issues.**

Big Data -> Useful Information

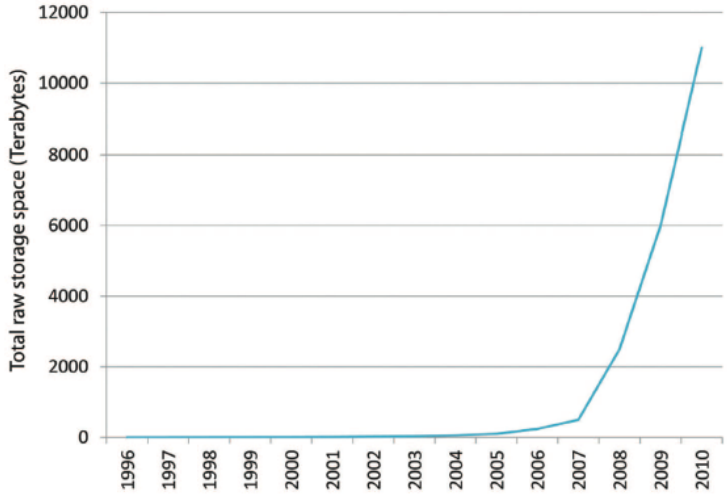


Figure 1. EMBL-EBI: Data storage increase [9].

Application	Data amount
Facebook	130 TB/day (user logs) 200-400 TB/day (pictures)
Google	25 PB/day (datasets)
Twitter	12 PB/day (datasets)
Large Hadron Collider (LHC)	60 TB/day (expected)
Large Synoptic Survey Telescope (LSST)	30 TB/day (expected)
Human Genome DNA	200 GB (per sequence)
X-ray image data	1TB/day
Cross-country (U.S.) Boeing 737	240 TB/flight
Mobile PC traffic	300 PB/month

Table I
DATA AMOUNT IN 2010 AND 2011.

SSD as a Storage Processing Element

Architecture Simulation - Parameters

CPU Power: use ITRS HP technology to evaluate dynamic and leakage power.
Number of Gates: 200M/core
Frequency: 2GHz.
Dynamic Power (per core): 5.04W
Leakage Power (per core): 0.340W

SSD Controller Power: use ITRS LOP technology to evaluate dynamic and leakage power.
Number of Gates: 20 millions per core (Assumption: 10% of the CPU).
Frequency: 1GHz.
Dynamic Power (per core): 0.156W.
Leakage Power (per core): 1.34mW.

Channel Processor Power: use ITRS LOP technology to evaluate dynamic and leakage power.
Number of Gates: 1K, 10K, 100K, 1M.
Frequency: 400MHz.
Dynamic Power (per core): 3.12uW, 31.2uW, 312uW, 3.12mW.
Leakage Power (per core): 67nW, 670nW , 6.7uW, 67uW.

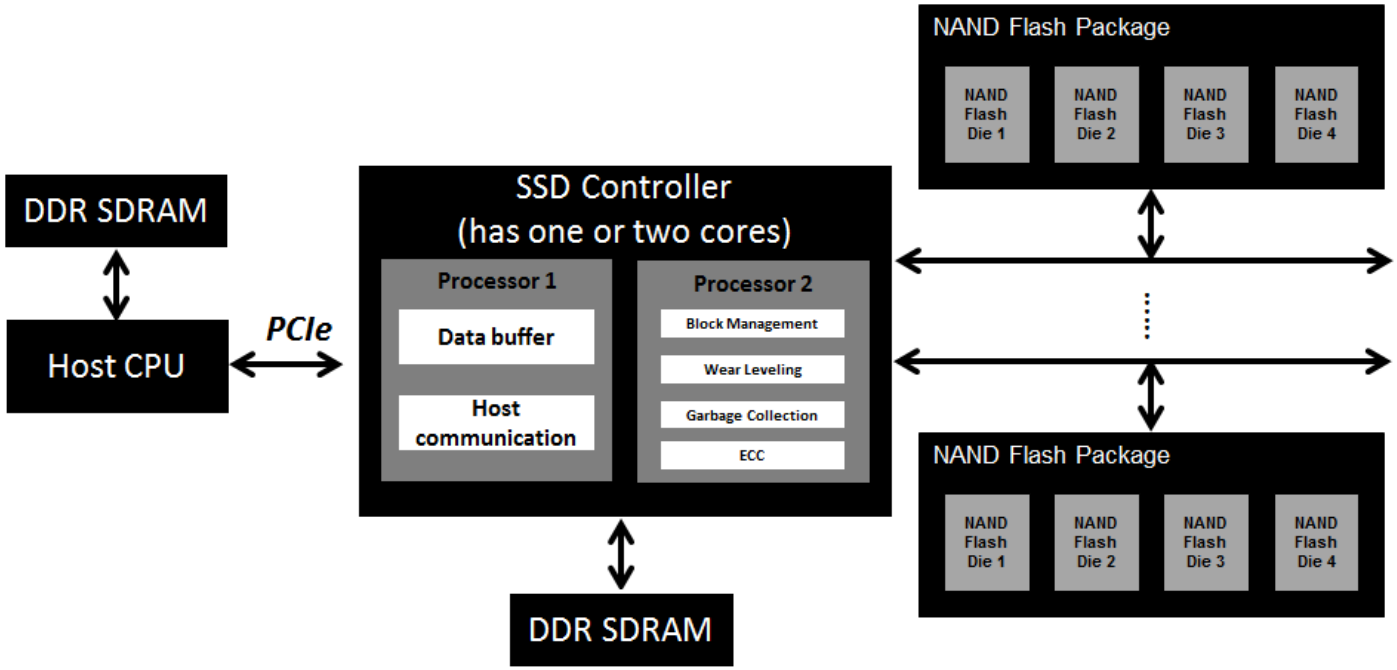
DDR SDRAM: use parameters from MICRON.
Dynamic Power (per 2GB): 438.3mW.
Leakage Power (per 2GB): 88.1mW.

NAND Flash: use parameters from MICRON.
Dynamic Power (per die): 0.04W.
Leakage Power (per die): 0.003W.

Host Interface: PCIe.
Dynamic Power (per GB): 37.5mW.
Leakage Power (per GB): 0.mW

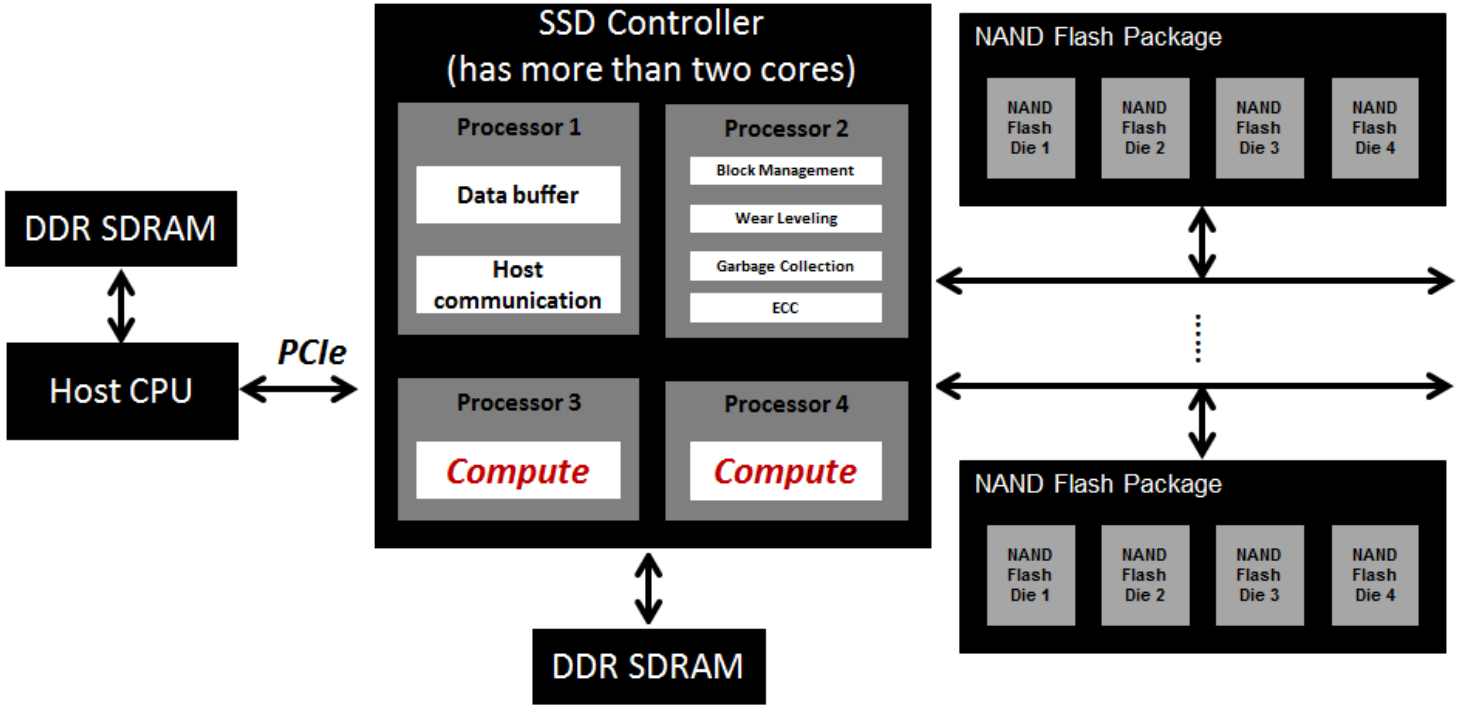
Architecture Block Diagrams

Baseline – SSD for Data, Compute in Host



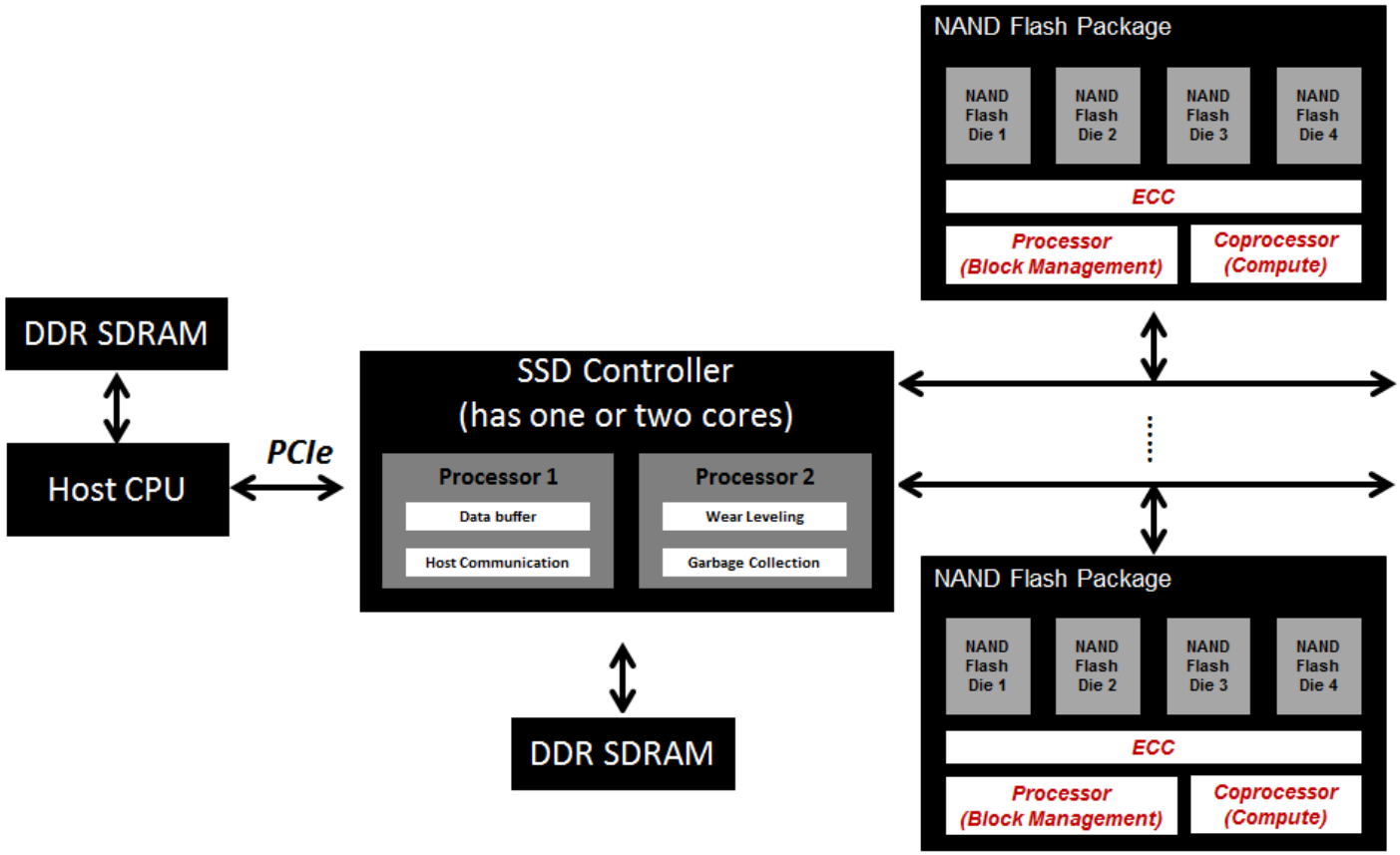
Architecture Block Diagrams

"Active Flash" – Compute in SSD Controller Processor



Architecture Block Diagrams

Storage Processing Element (SPE) – Compute in the Flash Channel Processors in each Flash package



Baseline Face Recognition

	1-Core	2-Core	4-Core	8-Core	16-Core
Average Processing Time of Facial Recognition Algorithm (ms)					
CPI = 100	52.7	26.4	13.3	6.80	3.50
CPI = 10	5.50	2.90	3.40	3.00	2.90
CPI = 1	3.00	2.80	2.80	2.70	2.70
CPI = 0.1	2.70	2.70	2.70	2.70	2.70
Average Power of Facial Recognition Algorithm (W)					
CPI = 100	294	289	286	287	286
CPI = 10	31.5	31.5	33.9	36.9	45.4
CPI = 1	6.36	7.12	9.17	12.6	20.1
CPI = 0.1	3.76	4.70	6.57	10.3	17.7
Average Energy of Facial Recognition Algorithm (mJ)					
CPI = 100	294	289	286	287	286
CPI = 10	31.5	31.5	33.9	36.9	45.4
CPI = 1	6.36	7.13	9.17	12.6	20.1
CPI = 0.1	3.76	4.71	6.57	10.3	17.7

Core = Host CPU Cores
CPI = clock cycles per instruction of single core in CPU

Active Flash Face Recognition

	1-Core	2-Core	4-Core	8-Core	16-Core
Average Processing Time of Facial Recognition Algorithm (ms)					
CPI = 100	52.6	26.4	13.3	6.70	3.40
CPI = 10	5.40	2.80	1.50	0.800	0.500
CPI = 1	0.700	0.400	0.300	0.300	0.300
Average Power of Facial Recognition Algorithm (W)					
CPI = 100	0.699	0.858	1.17	1.79	2.98
CPI = 10	0.716	0.881	1.18	1.70	2.48
CPI = 1	0.839	1.02	1.15	1.16	1.17
Average Energy of Facial Recognition Algorithm (mJ)					
CPI = 100	36.8	22.6	15.6	12.0	10.1
CPI = 10	3.86	2.47	1.78	1.36	1.24
CPI = 1	0.587	0.410	0.345	0.347	0.351

Core = SSD Controller Cores
CPI = clock cycles per instruction of single core in SSD controller

SPU Face Recognition

Channels	4	8	16	32
Average Processing Time (ms)				
Time	0.300	0.200	0.100	0.0500
Average Power of Facial Recognition Algorithm (W)				
Gates = 1K	0.887	1.23	1.87	2.98
Gates = 10K	0.887	1.23	1.87	2.98
Gates = 100K	0.888	1.23	1.88	2.99
Gates = 1M	0.899	1.26	1.92	3.06
Average Energy of Facial Recognition Algorithm (mJ)				
Gates = 1K	0.266	0.246	0.187	0.149
Gates = 10K	0.266	0.246	0.187	0.149
Gates = 100K	0.266	0.247	0.188	0.149
Gates = 1M	0.270	0.251	0.192	0.153

Baseline Boltzmann Machine

	1-Core	2-Core	4-Core	8-Core	16-Core
Average Processing Time of Boltzmann Machine Algorithm (ms)					
CPI = 100	0.952	0.497	0.270	0.157	0.0998
CPI = 10	0.134	0.0888	0.100	0.0943	0.0913
CPI = 1	0.0931	0.0907	0.0895	0.0889	0.0886
CPI = 0.1	0.0888	0.0886	0.0885	0.0884	0.0884
Average Power of Boltzmann Machine Algorithm (W)					
CPI = 100	5.20	10.0	18.7	33.3	55.4
CPI = 10	3.99	6.48	8.02	12.0	19.5
CPI = 1	1.43	2.37	4.21	7.90	15.3
CPI = 0.1	1.02	1.94	3.78	7.46	14.8
Average Energy of Boltzmann Machine Algorithm (mJ)					
CPI = 100	4.95	4.98	5.06	5.21	5.53
CPI = 10	0.534	0.576	0.804	1.13	1.78
CPI = 1	0.133	0.215	0.377	0.702	1.35
CPI = 0.1	0.0906	0.172	0.334	0.660	1.31

Core = Host CPU Cores
CPI = clock cycles per instruction of single core in CPU

Active Flash Boltzmann Machine

	1-Core	2-Core	4-Core	8-Core	16-Core
Average Processing Time of Facial Recognition Algorithm (ms)					
CPI = 100	34.2	17.1	8.59	4.31	2.18
CPI = 10	0.528	0.690	1.00	1.59	2.63
CPI = 1	0.381	0.210	0.124	0.0816	0.0602
Average Power of Facial Recognition Algorithm (W)					
CPI = 100	0.521	0.679	0.993	1.62	2.85
CPI = 10	0.528	0.690	1.00	1.59	2.63
CPI = 1	0.595	0.785	1.08	1.45	1.84
Average Energy of Facial Recognition Algorithm (mJ)					
CPI = 100	17.8	11.6	5.53	6.98	6.21
CPI = 10	1.83	1.21	0.897	0.742	0.665
CPI = 1	0.227	0.165	0.134	0.118	0.111

Core = SSD Controller Cores
CPI = clock cycles per instruction of single core in SSD controller

SPU Boltzmann Machine

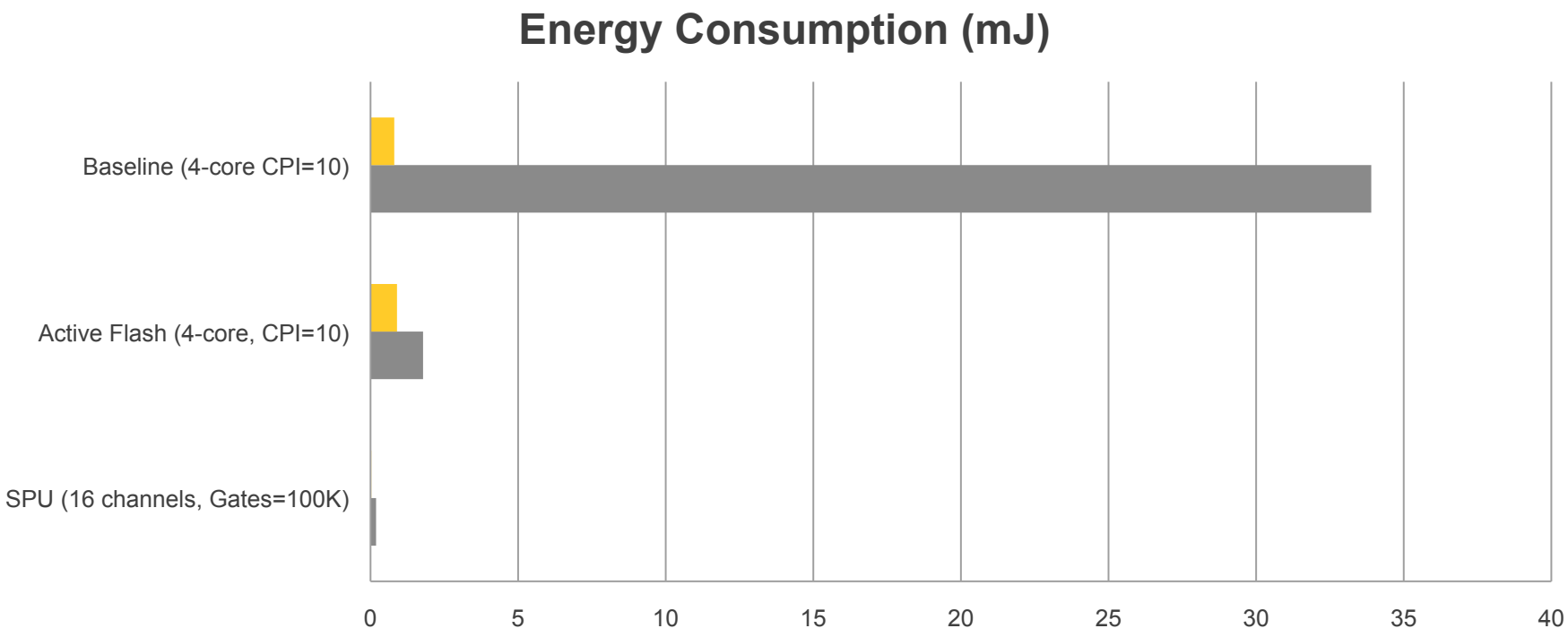
Channels	4	8	16	32
Time (ms)	0.0873	0.0584	0.0493	0.0482
Average Power of Facial Recognition Algorithm (W)				
Gates = 1K	0.415	0.451	0.486	0.512
Gates = 10K	0.415	0.451	0.486	0.512
Gates = 100K	0.415	0.452	0.488	0.514
Gates = 1M	0.424	0.465	0.503	0.529
Average Energy of Facial Recognition Algorithm (mJ)				
Gates = 1K	0.0362	0.0263	0.0240	0.0247
Gates = 10K	0.0362	0.0263	0.0240	0.0247
Gates = 100K	0.0363	0.0264	0.0240	0.0247
Gates = 1M	0.0370	0.0271	0.0248	0.0254

Summary

Facial recognition task – which is a proxy algorithm for content based image retrieval - Compute on 16 channel SSD is ~ 0.2mJ/face vs 30mJ/face computing on Host

Boltzmann machine task which is proxy for many machine learning and data intensive scientific compute algorithms – Compute on SSD is ~40X lower Joule/operation compared to Quad-Core host

Summary – cont'd



	SPU (16 channels, Gates=100K)	Active Flash (4-core, CPI=10)	Baseline (4-core CPI=10)
Boltzmann Machine	0.024	0.897	0.804
Facial Recognition	0.188	1.78	33.9

References

[1] “Hitting the memory wall: implications of the obvious”, WA Wulf, SA McKee - ACM SIGARCH computer architecture news, 1995

[2] “Reflections on the memory wall”, SA McKee - Proceedings of the 1st conference on Computing, 2004

[3] “Missing the memory wall: The case for processor/memory integration”, A Nowatzyk, F Pong, A Saulsbury - Architecture, 1996

[4] “Computing performance: Game over or next level?”, SH Fuller, LI Millett - Computer, 2011

[5] “Platform 2015: Intel processor and platform evolution for the next decade”, S Borkar, P Dubey, K Kahn, D Kuck, H Mulder

[6] “Dark silicon and the end of multicore scaling”, H Esmaeilzadeh, E Blem, RS Amant… (ISCA), 2011

[7] “GPUs and the future of parallel computing”, SW Keckler, WJ Dally, B Khailany, M Garland - Micro, 2011

[8] “The GPU computing era”, J Nickolls, WJ Dally - Micro, IEEE, 2010

[9] “Architecture at the End of Moore”, S Kaxiras - 2013 – Springer

[10] “The Shift to Cloud Computing: Forget the Technology, It’s About Economics”, Jim Cooke 2010

[11] “An Energy-Efficient Processor Architecture for Embedded Systems”, Balfour et. al. 2008

[12] “Trends in Computation, Communication and Storage and the Consequences for Data-intensive Science”, Oliveira, S.F., 2012

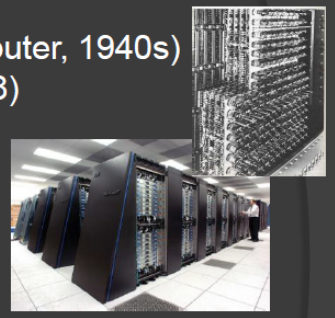
Thank You

What electronics and computers can learn from nature - Christoph Posch

Energy Efficiency

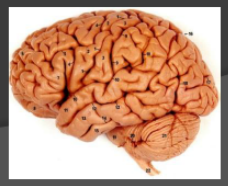
Progress of electronic information processing over past 60 years:

- dramatic improvements:
- from 5 Joules / instruction (vacuum tube computer, 1940s)
- to 0.0000000001 Joules / instruction (ARM968)
- 50,000,000,000 times better
- Raw performance increase about 1 million



Energy efficiency

- Chip: 10^{-11} J/operation
- Computer system level: 10^{-9} J/operation
- Brain: 10^{-15}J/operation
- Brain is 1 million times more energy efficient!!!



S. Furber, "The Dennis Gabor Lecture 2010: Building Brains" (2010)
C. Mead, "Neuromorphic Electronic Systems" (1990)

4

Bio-inspired Vision - and what electronics and computers can learn from nature
Christoph Posch - Austrian Institute of Technology AIT TWEPP 2011

Where is the Energy?

1 Million

- Cost of elementary operation – turning on transistor or activating a synapse – is about the same. (10^{-15}J)
- Lose a factor 100 because:
  - capacitance of gate is a small fraction of capacitance of the node
  - spend most energy charging up wires
- Use many transistors to do one operation (typically switch 10000).
  - information encoding: "0", "1"
  - elementary logic operations (AND, OR, NOT)

C. Mead: "We pay a *factor 10000* in energy for taking out the beautiful physics from the transistor, mash it up into "0"and "1" and then painfully building it back up with gates and operations to reinvent [e.g.] the multiplication …"

C. Mead, "Neuromorphic Electronic Systems" Proc. IEEE, (1990)

5

Bio-inspired Vision - and what electronics and computers can learn from nature
Christoph Posch - Austrian Institute of Technology AIT TWEPP 2011

Brain vs. Computer - II

At the system level, brains are at least **1 million times more power efficient** than computers. **Why?**

Cost of elementary operation (turning on transistor or activating synapse) is **about the same**. It's not some magic about physics. (10^{-15} J)

Computer	Brain
Fast global clock	Self-timed, data driven
Bit-perfect deterministic logical state	Synapses are stochastic! Computation dances digital→analog→digital
Memory distant to computation	Synaptic memory at computation
Fast, high resolution, constant sample rate analog-to-digital converters	Low resolution adaptive data-driven quantizers (spiking neurons)

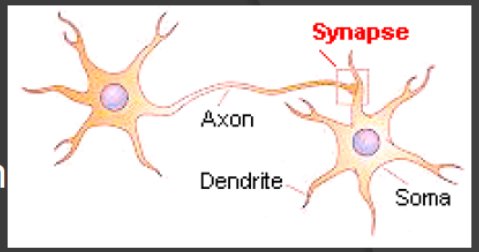
Mobility of electrons in silicon is about **10^7 times** that of ions in solution.

T. Delbruck, "Spiking silicon retina for digital vision". IEEE DLP lecture

36

Bio-inspired Vision - and what electronics and computers can learn from nature
Christoph Posch - Austrian Institute of Technology AIT TWEPP 2011

Processing and Storage



- N1 **spikes**—pulse travels down the axon to the synapse of target N2.
- The synapse of N2—having **stored its own state locally**—evaluates the importance of the information coming from N1 by integrating it with **own previous state** and **strength** of connection to N1
- **Two pieces of information**—signal from N1 and state of N2's synapse—**flow** toward body of N2
- When information reaches N2, there is only a **single value—all processing has already taken place during the information transfer**.
- Storage and processing happen at the ***same time*** and in the ***same place***.
- This **LOCALITY** is one of main reasons for **energy efficiency** of biological brains

53

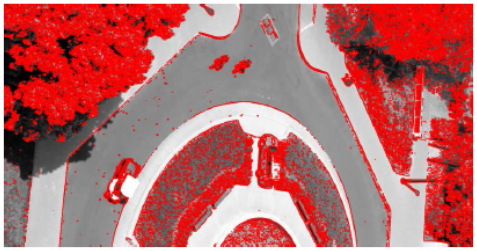
Bio-inspired Vision - and what electronics and computers can learn from nature
Christoph Posch - Austrian Institute of Technology AIT TWEPP 2011

Dan Hammerstrom - DARPA

Solution: A New Computational Model

New Paradigm – Non-Boolean, Probabilistic Computing

1. Computing occurs by the physics of the devices (highly parallel)
2. Devices perform the computational equivalent of hundreds of discrete digital operations
3. The model can be configured into hierarchies that accomplish most of the computational work required by the application



Sensor Data: Active Edges located (in red)

Example: Find Features in Sensor Data (7x7 Gabor Edge Finding, 10 Giga-pixel Array)

Boolean Computation

- Processor: Intel 6 Core i7, GOPS: 6.7
- 1 inference is 140 operations/kernel, 24 kernels are compared / pixel
- GOPs/watt: 0.1
- Compute time = 7,700 sec
- **460 kilo-joules** (60 watts for 7700 seconds)

Analog Direct Device Computation

- Processor: 10 X 10 Array of coupled oscillators Giga-Inferences/sec = 400 (56k GOPS equivalent)
- Compute time = 0.04 sec
- **430 milli-joules**