Computing's Energy Problem:

(and what we can do about it)

Mark Horowitz

Stanford University horowitz@ee.stanford.edu

Everything Has A Computer Inside















International Solid-State Circuits Conference

The Reason is Simple: Moore's Law Made Gates Cheap

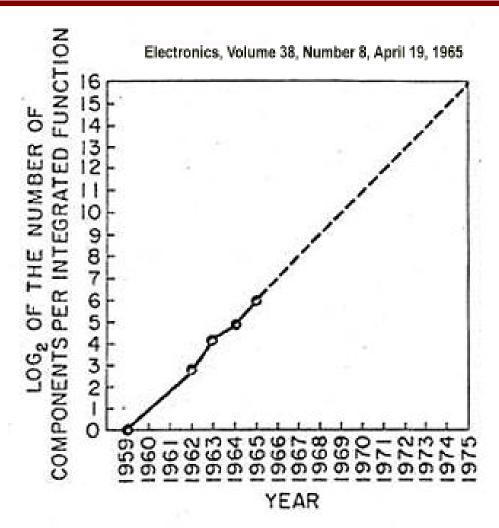
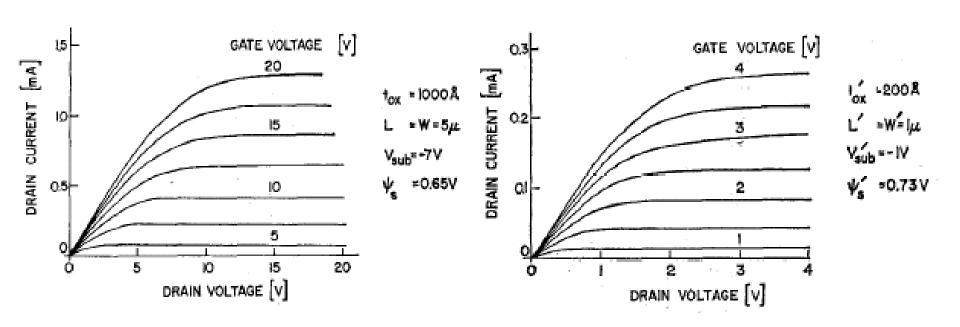


Fig. 2 Number of components per Integrated function for minimum cost per component extrapolated vs time.

Dennard's Scaling Made Them Fast & Low Energy



The triple play:

- Get more gates,
- · Gates get faster,
- Energy per switch

1/L ²	$1/\alpha^2$
· / —	., 0

CV/i α

 CV^2 α^3

Dennard, JSSC, pp. 256-268, Oct. 1974

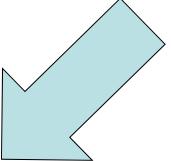
Our Expectation

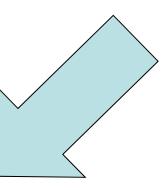
Cray-1: world's fastest computer 1976-1982

- 64Mb memory (50ns cycle time)
- 40Kb register (6ns cycle time)
- ~1 million gates (4/5 input NAND)
- 80MHz clock
- 115kW

In 45nm (30 years later)

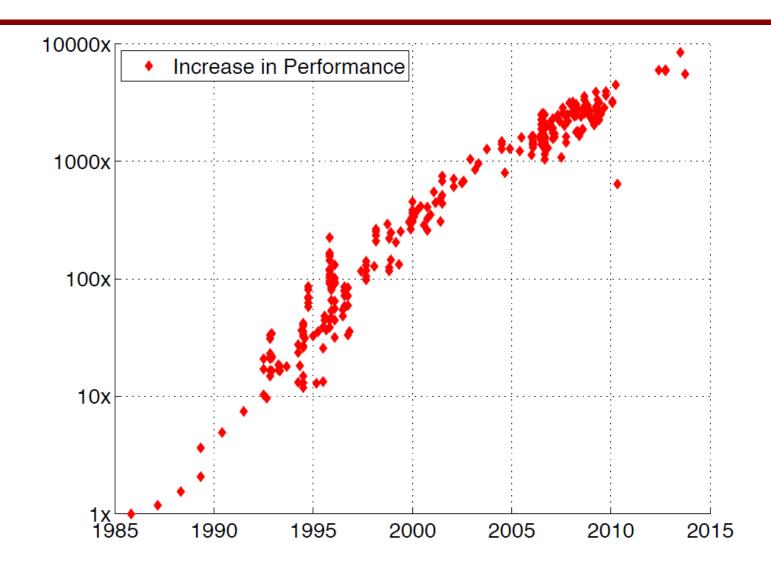
- $< 3 \text{ mm}^2$
- > 1 GHz
- ~ 1 W



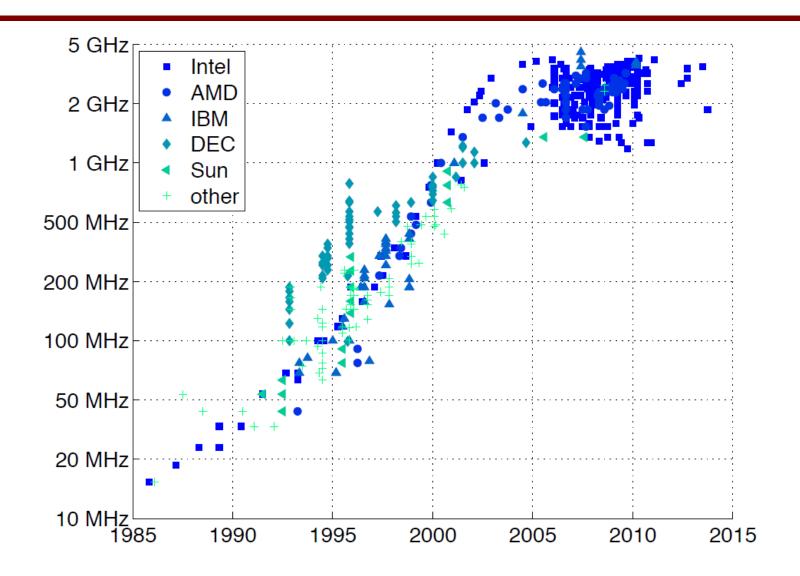




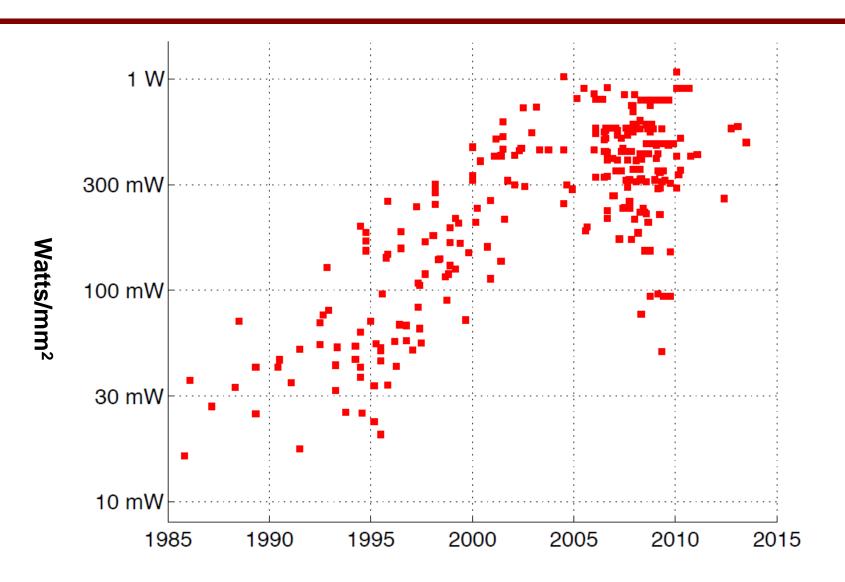
Supporting Evidence



Houston, We Have A Problem

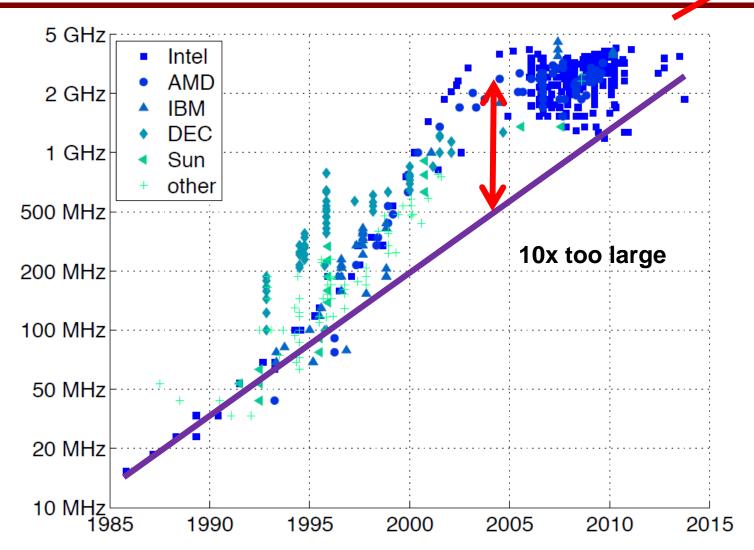


The Power Limit

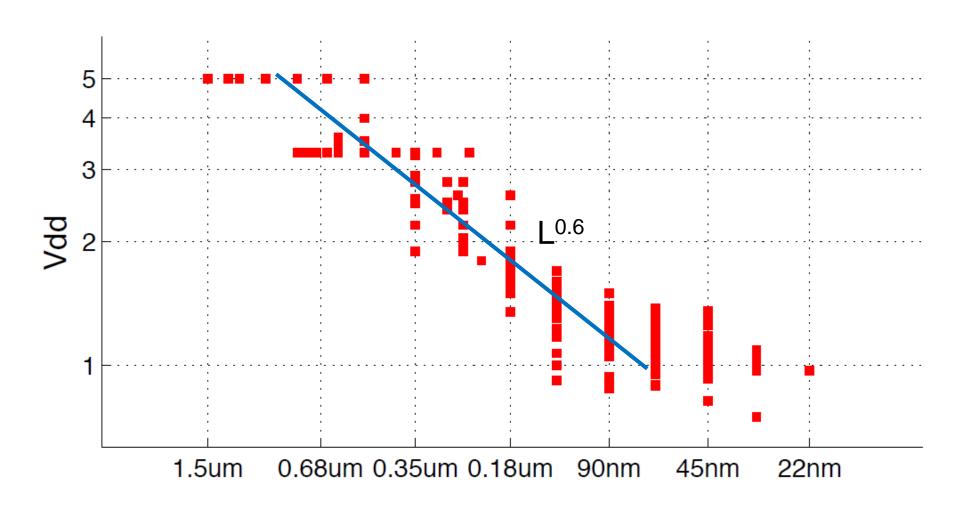


Clever

Power Increased Because We Were Greedy

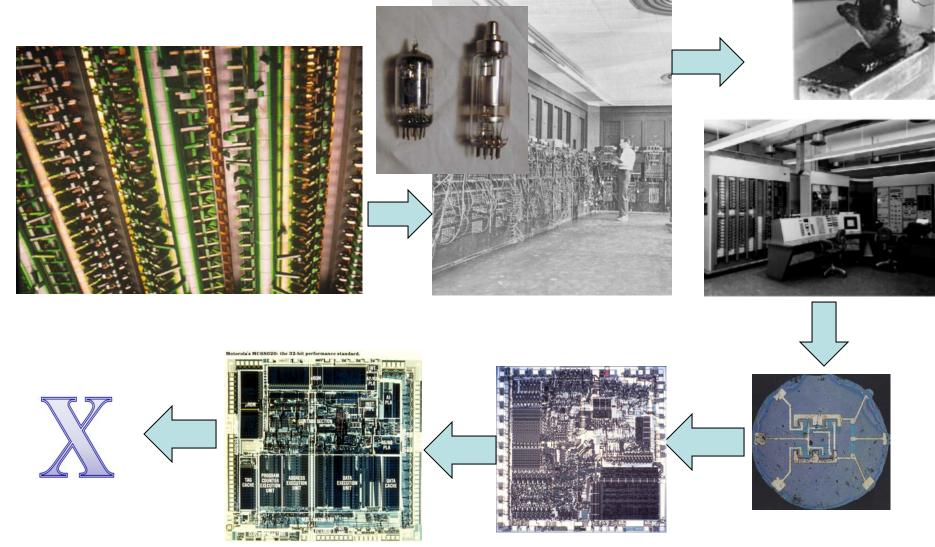


This Power Problem Is Not Going Away: $P = \alpha C * Vdd^2 * f$



Think About It

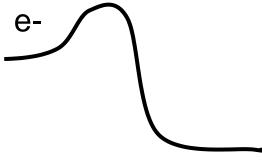
Technology to the Rescue?



Problems w/ Replacing CMOS

Pretty fundamental physics

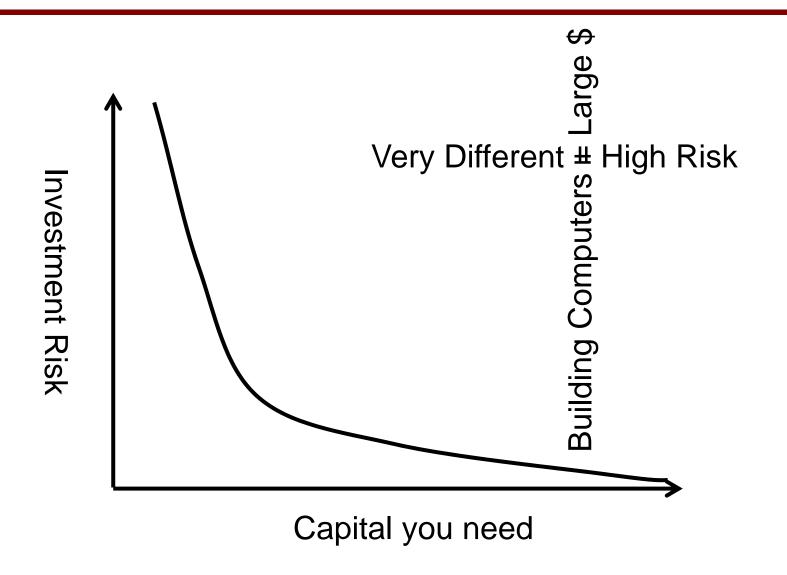
Avoiding this problem will be hard



Its capability is pretty amazing

• fJ/gate, 10ps delays, 10⁹ working devices

Catch - 22



The Truth About Innovation



Start by creating new markets

1.1: Computing's Energy Problem: (and what we can do about it)

Our CMOS Future

Will see tremendous innovative uses of computation

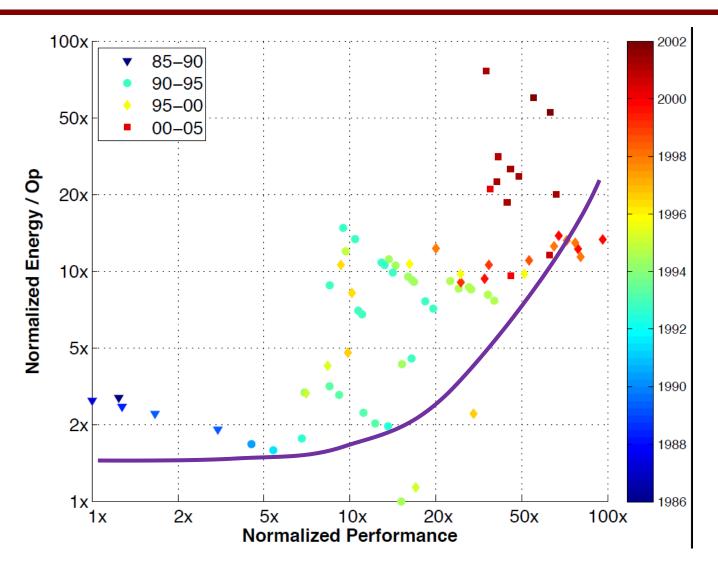
- Capability of today's technology is incredible
- Can add computing and communication for nearly \$0
- Key questions are what problems need to be solved?

Most performance system will be energy limited

These systems will be optimized for energy efficiency

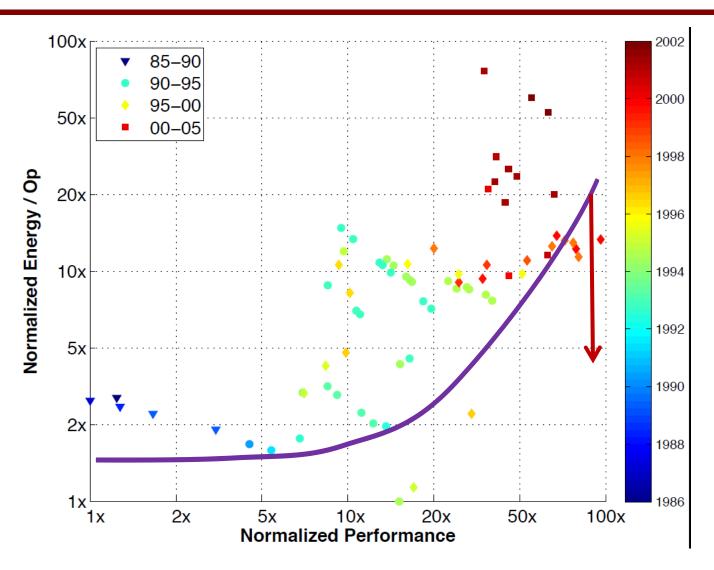
Power = Energy/Op * Ops/sec

Processor Energy – Delay Trade-off



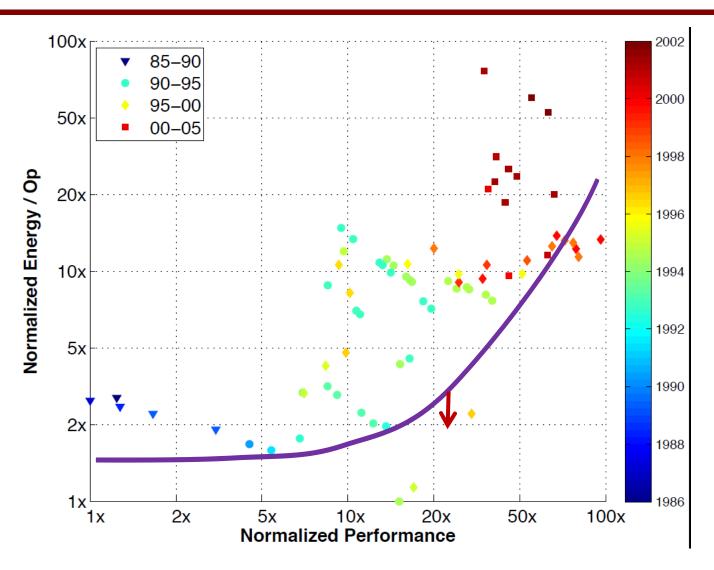
http://cpudb.stanford.edu/

The Rise of Multi-Core Processors



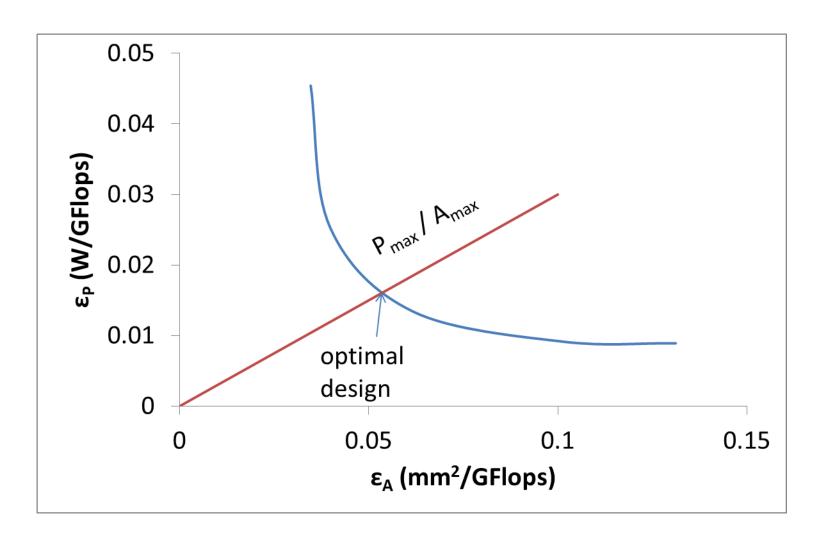
http://cpudb.stanford.edu/

The Stagnation of Multi-Core Processors



http://cpudb.stanford.edu/

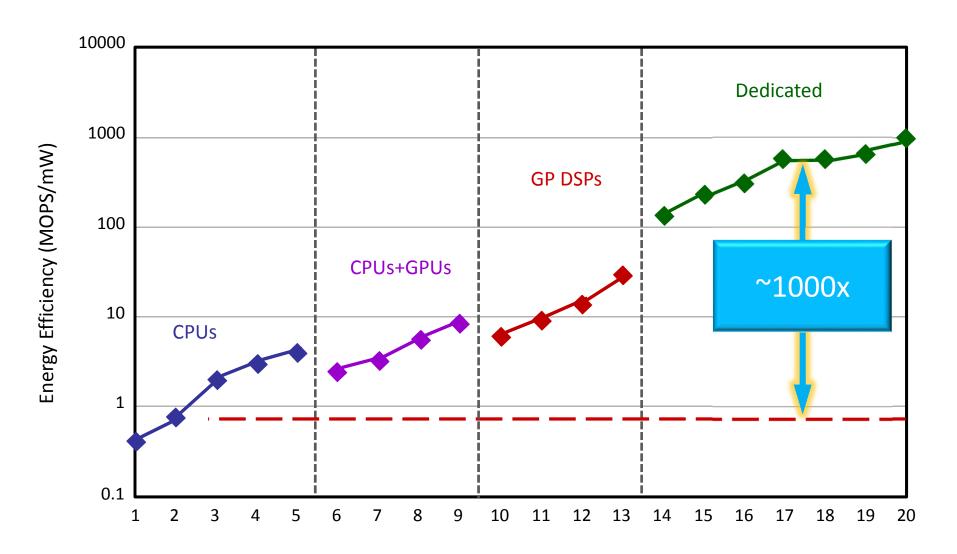
Optimizing Parallel Machines (GPUs)



Have A Shiny Ball, Now What?



Signal Processing ASICs



The Push For Specialized Hardware

Dark Silicon and the End of Multic

Hadi Esmaeilzadeh Emily Blem Renée St. Amante *University of Washington hadianeh@cs.washington.edu blem@cs.wisc.edu se The University of Texas at

ABSTRACT

Since 2005, processor designers have increase ploit Moore's Law scaling, rather than focusing from mone 3 cames of Dennard scaling, to which ticore parts is partially a response, may soon limit mu just as single-core scaling has been curtailed. This p. multicore scaling limits by combining device scaling, a scaling, and multicore scaling to measure the speedup pole a set of parallel workloads for the next five technology gener, For device scaling, we use both the ITRS projections and of more conservative device scaling parameters. To model sing core scaling, we combine measurements from over 150 processor. to derive Pareto-optimal frontiers for area/performance and powexperiormance. Finally, to model multicore scaling, we build a detailed performance model of upper-bound performance and lowerbound core power. The multicore designs we study include singlethreaded CPU-like and massively threaded GPU-like multicore chip organizations with symmetric, asymmetric, dynamic, and composed organizations with symmetric, any minimum, and composed topologies. The study shows that regardless of chip organization and topology, multicore scaling is power limited to a degree nor widely appreciated by the computing community. Even at 22 nm (just one year from now), 21% of a fixed-size chip must be powered off, and at 8 nm, this number grows to more than 50%. Through 2024, only 7.9x average speedup is possible across commonly used parallel workloads, leaving a nearly 24-fold gap from a target of doubled performance per generation.

Categories and Subject Descriptors: C.0 [Computer Systems Organization] General — Modeling of computer architecture; Co [Computer Systems Organization] General — System architectures General Terms; Design, Measurement, Performance Keywords: Dark Silicon, Modeling, Power, Technology Scalin

Conservations Cores: Computations the Energy of Mature Sammer Saila Taylor Michael Bedford Taylor . University of California, San Diego

University of California, San Diego

University of California, San Diego

Evenkatesh jeampson, negouldin, sat vobryksin juugomar, avvanson, mbtaylor) @cs. ucsd edu Prover Consequently the rate at which we can switch francistors to disciplate the heat created by those fix for our ability to disciplate the heat created by those transistors. ansistors.

The result is a technology imposed utilization wall that lime.

The result is a technology can use at full speed at one time.

The fraction of the chip we can use at full speed. The result is a technology-imposed utilization walt that limits.

The result is a technology-imposed utilization walt that limits inc.

the traction of the chip we can TSMC process show that we can

Our experiments with a 45 nm TSMC process show that we the fraction of the chip we can use at full speed at one time.

Our experiments with a 45 nm 75MC process show that we can

our experiments with a 45 nm 75MC process are full frequency within an

switch less than 756 of a 300mm. Ganesh Venkalesh Our experiments with a 45 nm TSMC process show that within an analysis of a 300mm of the frequency within a 300mm of the arms and CMOS scaling switch less than 7% of a 300mm brokerion switch less than 7% of a 300mm brokerion switch less than 7% of a 300mm brokerion switch les Switch less than 7% of a 300mm² die af tuit frequency within an angle of a 300mm² die af tuit frequency within an angle of a 300mm² die af tuit frequency within and constraints that this percentage will decrease to less than 3.5% theory suggests that this percentage 80W power budget, TTRS roadmap projections and cMOS scaling to less than 3.3% will be crease to less than 3.3% with the crease to less than 3.3% with the crease to less than 3.3% with the crease to less than the crease to less than the continue to decrease to almost half with each theory suggests that the continue to decrease to almost half with continue to decrease to almost half with continue to decrease the almost half with continue to decrease the almost half with case the continue to decrease the almost half with case the continue to decrease the almost half with case the continue to decrease the almost half with case the continue to decrease the almost half with case the continue to decrease the almost half with case the continue to decrease the almost half with case the continue to decrease the almost half with case the continue to decrease the almost half with case the continue to decrease the almost half with the continue to decrease the continue the continue to decrease the continue to decrease the continue th Vladyslav Bryksin theory suggests that this Percentage will decrease to less than 3.5% among half with each theory suggests that this Percentage will decrease to less than 3.5% among half with each theory suggests that this Percentage will decrease to less than 3.5% among the less than 3. in 32 nm, and will comfine to decrease by almost half will not and even turner with 3.17 integration. and even turner with 3.27 inceptation. and even turner with 3.27 inceptation. The utilization wall are already indirect. The effects of the utilization wall are already indirect. rocuss generation—and even further with 3-D integration.

and even further with 3-D integration.

The effects of the utilization wall are already indirectly approximately and the provides a "turbo modern provides a "turbo mem in modern processors; Intel's Nebalem provides a "turbo mem in modern processors; Intel's Nebalem provides a "turbo mem in modern processors; Intel's Nebalem provides a "turbo mem in modern processors; Intel's Nebalem provides a "turbo memory in modern processors; Intel's Nebalem provides a "turbo memory in modern processors; Intel's Nebalem provides a "turbo memory in modern processors; Intel's Nebalem provides a "turbo memory in modern processors; Intel's Nebalem provides a "turbo memory in modern processors; Intel's Nebalem provides a "turbo memory in modern processors; Intel's Nebalem provides a "turbo memory in modern processors; Intel's Nebalem provides a "turbo memory in modern processors; Intel's Nebalem provides a "turbo memory in modern processors; Intel's Nebalem provides a "turbo memory in modern processors; Intel's Nebalem processors; Intel's The effects of the utilization wall are already indirectly appropriate a tripper page in modern processors in order to run others at tripper pagen, that powers our some comes in order to run others at tripper mode, that powers our some comes in order to run others at tripper mode. parent in modern processors: Intel's Nebalem provides a "turbo the house in modern processors in order to run others at higher transmoder, that powers our some indication is that even though native transmoder. Another strong indication is that even though native transmoder. Another strong indication is that even though native transmoder. AUSITACE

Growing transistor counts, limited power budgets, and the break.

Growing transistor counts are currently consolring to create a military consolring to create a military consolring transistor counts. mode" that powers off some cones in order to run others at trighter trans-speeds. A nother strong indication is that given though native trans-speeds. A nother strong indication is that given though a process Growing transistor counts, limited power budgets, and the break.

Growing transistor counts, limited power budgets, and the amilitar.

Growing transistor counts, limited power budgets, and the amilitar.

Growing transistor counts, limited power budgets, and the break.

Growing transistor counts, limited power budgets, and the break.

Growing transistor counts, limited power budgets, and the break. down of voltage scaling are currently conspiring to create a military constituent of a chip that can run at full speed to create a military constituent of a chip that can run at full speed of a chip that can run at full speed on the constituent of a chip that can run at full speed on the constituent of a chip that can run at full speed on the constituent of the constituen tion wall that finites the frection of a chir final can run at full speed a chir final can run at full speed to the first the frection of a chir final can run at full speed to the run of the first the frection from the first t

octime a form of targeted reconfigurability, that allows

new versions of the software they target. Our re-

ation cores can reduce energy consumption

and on the recent lifetime of individual

at one time, in this regime, specialtized, energy-efficient processors to execute under the arminerase parallelism by reducing the Per-Computations to execute under the can increase parallelism by more computations to execute under the quirenests and allowing more computations. can increase Parallelism by reducing the per-computation private the quirements and allowing more computations to execute under the quirements and allowing more computations to paper introduces consumer that allowing more computations to execute under the paper introduces consumer to pursue this goal, this paper introduces consumer to pursue this goal, this paper introduces consumer to pursue this goal, this paper introduces consumer to pursue this goal. quirenents and allowing more computations to execute under the programments and allowing more computations to execute under the programments and allowing more computations to execute under the programments are specialized programments are programments. To pursue this goal, this paper introduced programments are specialized programments and allowing more cores, or occores, are specialized programments are programments. Same tower hudget. To pursue this goal, this paper introduces continued pro-servation cores. Conservation cores, or c-cores, and energy delay instead or servation cores on reducing onergy and energy delay instead or egsoors that rocus on reducing onergy. servation cores. Conservation cores, or c-cores are specialized pro-servation cores. Conservation cores, or c-cores are specialized pro-cessors that focus on reducing cores, or c-cores are specialized pro-servation cores. Conservation cores, or c-cores are specialized pro-cessors that focus on reducing cores, or c-cores are specialized pro-cessors that focus on reducing cores, or c-cores are specialized pro-ting cores. Conservation cores, or c-cores are specialized pro-increasing pro-page cores. cessors that focus on reducine energy and energy makes c-cores an ex-trus focus on reducine energy makes c-cores andidates. This focus on energy makes coor candidates increasing performance. This focus on energy makes focus on energy makes cover and dates. increasing performance. This focus on energy makes ecopes an ex-increasing performance. This focus on energy makes ecopes. We original increasing performance applications that would be poor candidates. cellent match for many applications that would be poor candidates. We present for hardware acceleration (e.g., irregular integer cores from application for hardware acceleration (e.g., irregular integers from application at the property of the property o for hardware acceleration (e.g., irregular integer codes). We present for hardware acceleration (e.g., irregular integer codes). We present for hardware acceleration (e.g., irregular integer codes). We present for hardware acceleration (e.g., irregular integer codes). We present for hardware acceleration (e.g., irregular integer codes). We present for hardware acceleration (e.g., irregular integer codes). We present for hardware acceleration (e.g., irregular integer codes). We present for hardware acceleration (e.g., irregular integer codes). We present for hardware acceleration (e.g., irregular integer codes). We present for hardware acceleration (e.g., irregular integer codes). We present for hardware acceleration (e.g., irregular integer codes). We present for hardware acceleration (e.g., irregular integer codes). a toolchain for automatically synthesizine cooks from application produce on source code and demonstrate that they can significantly. The occopes and demonstrate that they can applications. The occopes source code and demonstrate that they can significantly reduce entrance and demonstrate that they can significantly reduce entrance and allows allows on a wide range of applications. The allows are and energy delay for a wide range of a policy and energy delay for a wide range of a policy and energy delay for a wide range of a policy and energy delay for a wide range of a policy and energy delay for a wide range of a policy of a policy of a wide range of a policy of a policy of a wide range of a policy of parallelism nd energy delay for a wide range of applications. The e-cores chip, it is cn will be in the lo. performance of p of core doubling? Such a study mus chip organizations, and and power limits at eac. ers all those factors togeth

efit. S

integrat

ing doma.

In this regime, reducing percoperation onergy [19] translate the system. If the system is the system of the system is the conditional for the system of the conditional for the system of the conditional for the system of the sy directly into increased potential parallelism for the system if a second potential parallelism for the system is normal to consume less powerful the second potential to consume less powerful in normal to consume the new in normal transfer of the new in normal transfer of the new in the new interpretations and the new interpretations and the second transfer of the seco given computation can be made to consume less power at the same computations can be run in parties of performance, other bushess Whom violating the proper budget will with conservation to the paper attacks the utilization arealization arealization conservation cores. Or economic and arealization cores. Conservation cores, or c-cores, are application-specific by the control cores, or c-cores, are application specific consumers of network consumers of network consumers of the purpose of network circums of the control computationally intensive applications. never or percompance, other company, acres or percompany, other budget, without violating the power maintaining, with the maintaining of the perconagnet of the perco Circuits created for the purpose of reducing created is no low Computationally intensive of trail resources at cases is employed that to care the nature where of trail resources at cases Computationally intensive applications. Since it is no lor ble to run the online chip at full frequency at once, it may

Speeds. A nother strong indication is that even though native translated such a speed substantially speeds. A nother strong indication is that even though native from the speed substantially speeds. A nother strong indication is that even the strong indication is that even though native translation is that even the strong is the strong in the strong in the strong is the strong in the strong is the strong in the strong in the strong is the strong in the strong in the strong is the strong in the st

ever the last 5 years.

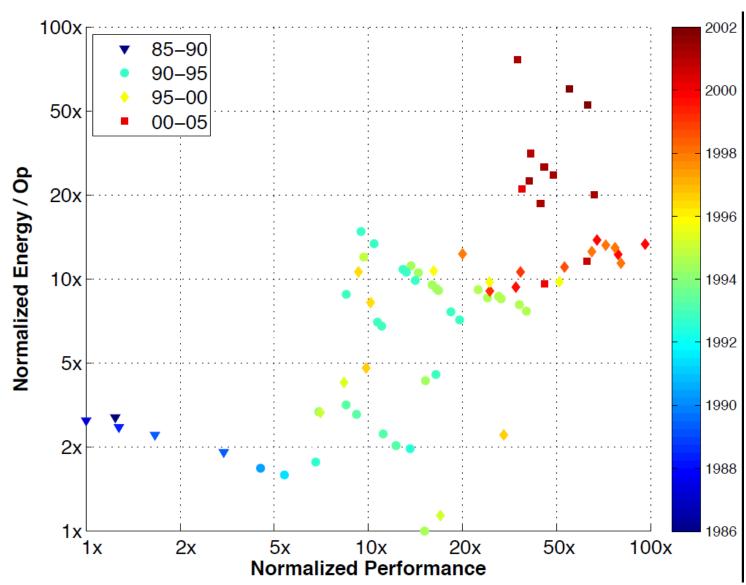
sissor switching speeds have continued to doubte every two process and increased substantially generations, processor frequencies have not increased substantially generations, processor frequencies have not increased substantially generations.

wer the last's years, reducing per-operation energy 1191 translates for the system. If the free the system is the system for the system from the free three free three t

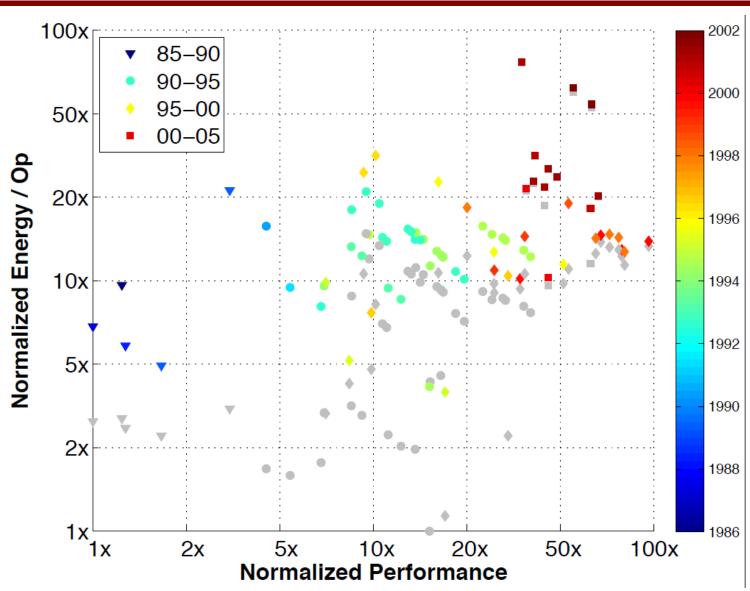
Before Talking About Specialization



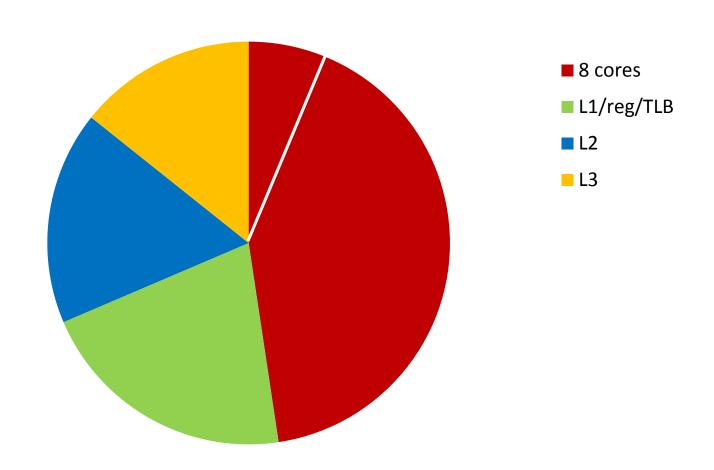
Don't Forget Memory System Energy



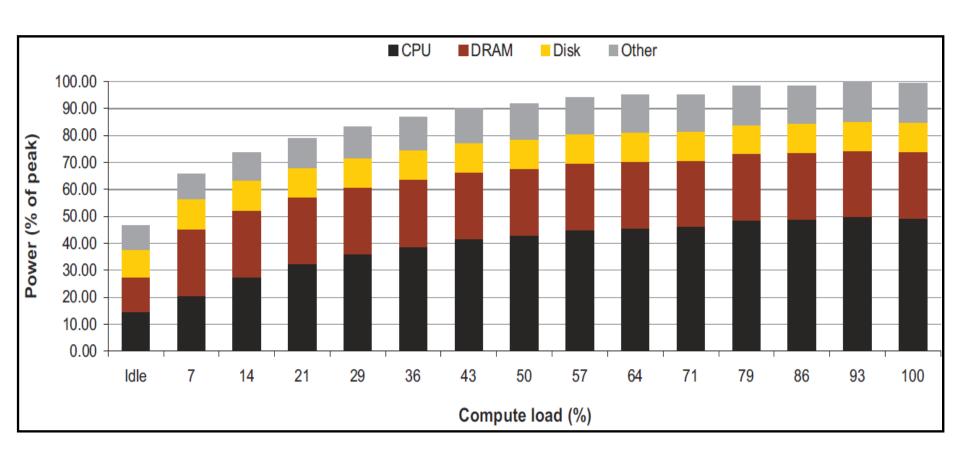
Processor Energy w/ Corrected Cache Sizes



Processor Energy Breakdown

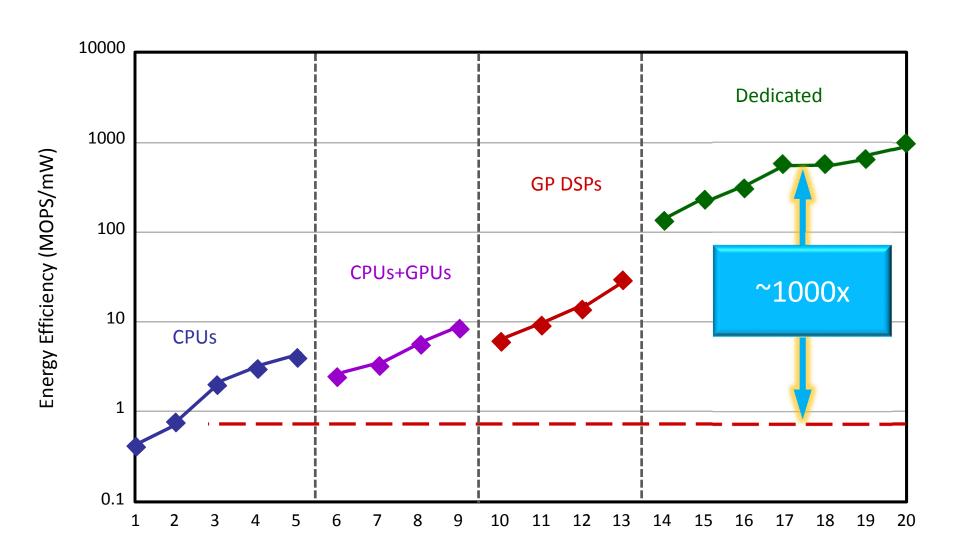


Data Center Energy Specs



SO HOW WILL ACCELERATORS HELP?

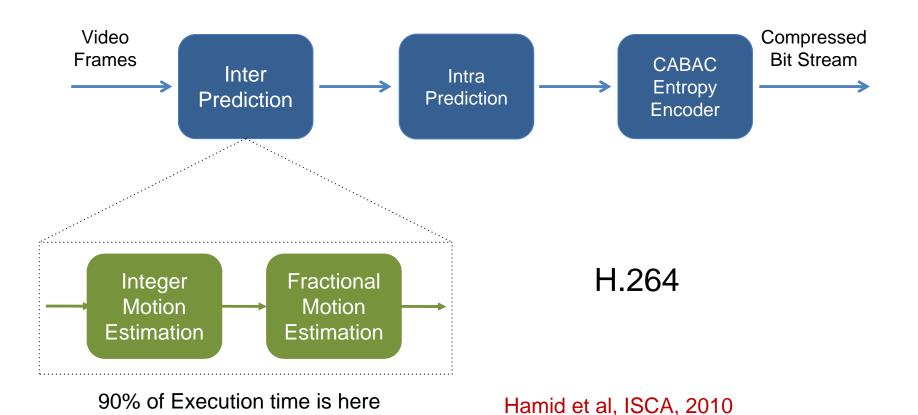
What Is Going On Here?



ASIC's Dirty Little Secret

All the ASIC applications have absurd locality

And work on short integer data



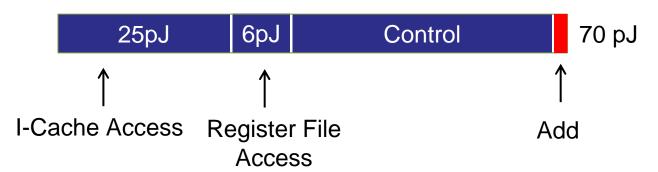
Rough Energy Numbers (45nm)

Integer	
Add	
8 bit	0.03pJ
32 bit	0.1pJ
Mult	
8 bit	0.2pJ
32 bit	3 pJ

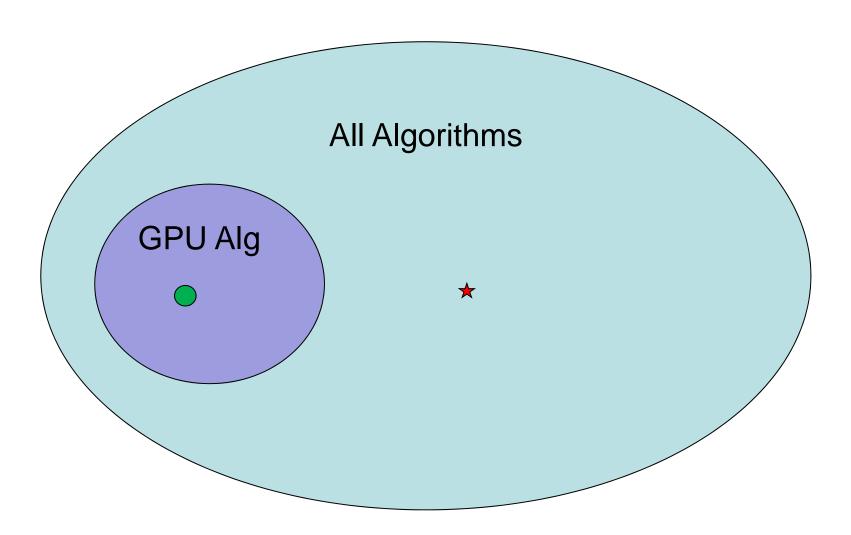
FP	
FAdd	
16 bit	0.4pJ
32 bit	0.9pJ
FMult	
16 bit	1pJ
32 bit	4pJ

Memory	
Cache	(64bit)
8KB	10pJ
32KB	20pJ
1MB	100pJ
DRAM	1.3-2.6nJ

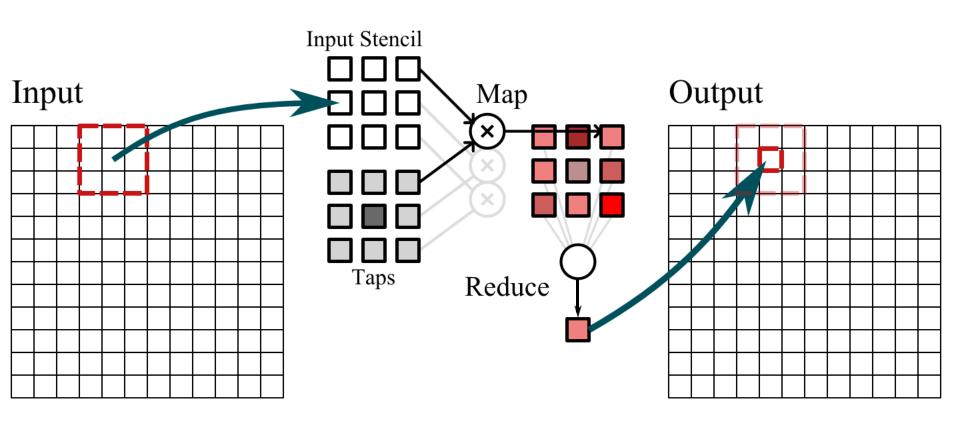
Instruction Energy Breakdown

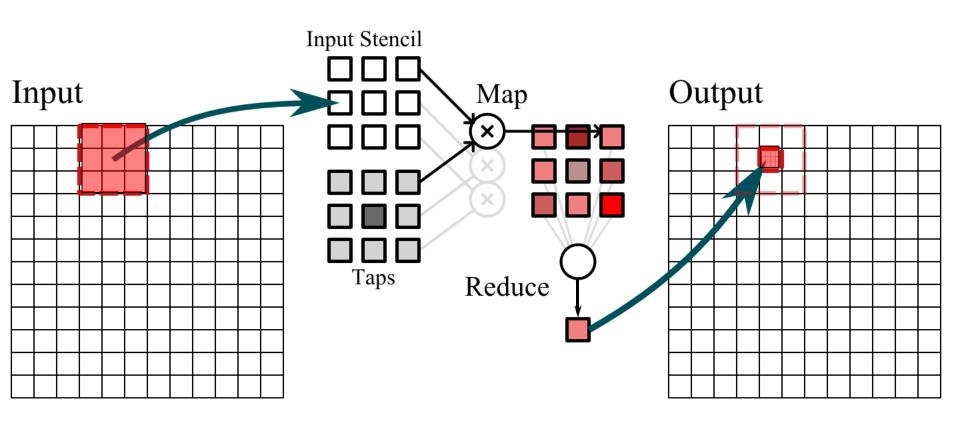


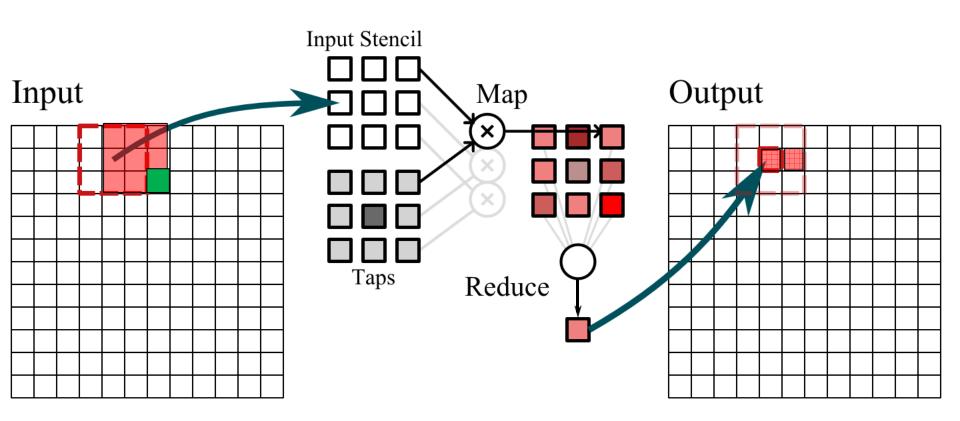
The Truth: It's More About the Algorithm then the Hardware

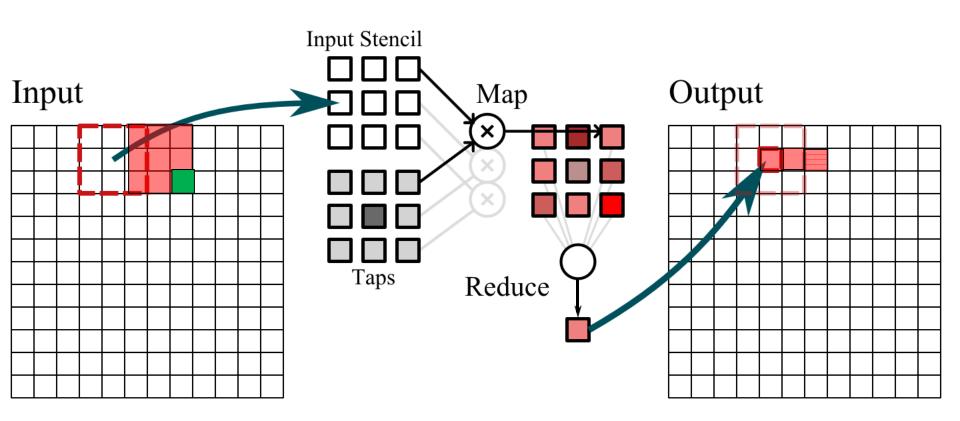




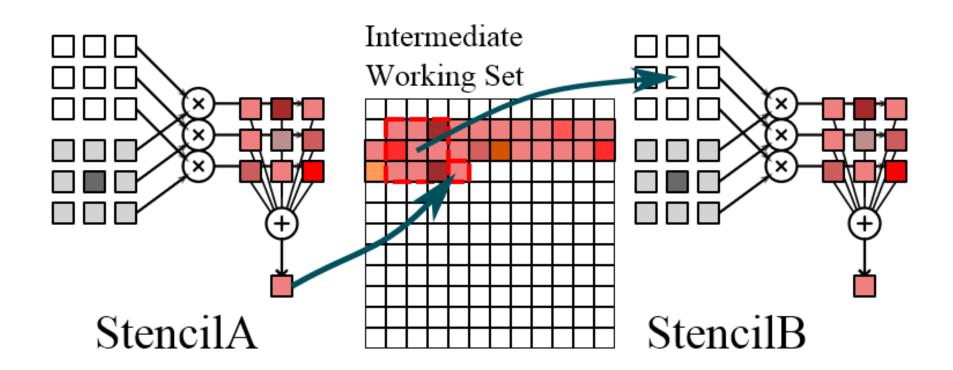








Compose These Cores into a Pipeline



Program in space, not time

Makes building programmable hardware more difficult

Working on System to Explore This Space

Takes high-level program

Graph of stencil kernels

Maps to hardware level assembly

Compute graph of operations for each kernel

Currently we map the result to:

FPGA, custom ASIC

Enabling Innovation

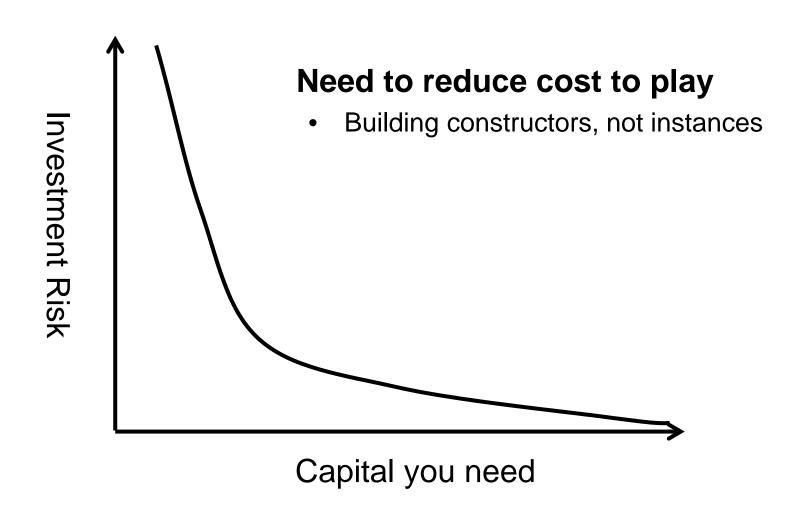
You don't just compile applications to efficiency

Need to tweak the application to fit constraints

Need to enable application experts to play

They know how to "cheat" and still get good results

Remember This Trade-off?



Not All Systems Are On The Bleeding Edge



App Store For Hardware



There's almost no limit to what iPhone can do.

The App Store has the best selection of mobile apps — from Apple and third-party developers. And they're all designed specifically for iPhone. The more apps you download, the more you'll realize your iPhone can do just about anything you can imagine.



Challenge



What Arduino can do

Arduino can sense the environment by receiving input from a variety of sensors and can affect its surroundings by controlling lights, motors, and other actuators. The microcontroller on the board is programmed using the Arduino programming language (based on Wiring) and the Arduino development environment

Community

The community of Arduino enthusiasts is vast, and includes region specific groups and special interest groups. The community is an excellent further source of assistance on all topics such as accessory selection, project assistance, and ideas of all sorts.

A New Hope

If technology is scaling more slowly

- We can incorporate current design knowledge into tools
- To create extensible system constructors

If killer products are going to be application driven

Application experts need to design them

We can leverage the 1st bullet to enable the 2nd

To usher in a new wave of innovative computing products