INVITED PLENARY TALK FOR VLSI TECHNOLOGY SYMPOSIUM 2011

TECHNOLOGY IMPACTS FROM THE NEW WAVE OF ARCHITECTURES FOR MEDIA-RICH WORKLOADS

Samuel Naffziger
AMD Corporate Fellow

August 26th, 2011 (Original presentation June 14th , 2011)

Outline

- Introduction
- The new workloads and demands on computation
- Characteristics of serial and parallel computation
- The Accelerated Processing Unit (APU) architecture
- APU architecture implications for technology
- Summary

The Big Experience/Small Form Factor Paradox

Technology	Mid 1990s	Mid 2000s	Now: Parallel/Data-Dense
Display	4:3 @ 0.5 megapixel	4:3 @ 1.2 megapixels	16:9 @ 7 megapixels
Content	Email, film & scanners	Digital cameras, SD webcams (1-5 MB files)	HD video flipcams, phones, webcams (1GB)
Online	Text and low res photos	WWW and streaming SD video	3D Internet apps and HD video online, social networking w/HD files
Multimedia	CD-ROM	DVDs	3D Blu-ray HD
Interface	Mouse & keyboard	Mouse & keyboard	Multi-touch, facial/gesture/voice recognition + mouse & keyboard
Battery Life*	1-2 Hours	3-4 Hours	All day computing (8+ Hours)

Form Factors

Early Internet and Multimedia Experiences

Standard-definition Internet

Immersive and interactive performance

Workloads

*Resting battery life as measured with industry standard tests.

Focusing on the experiences that matter

Consumer PC Usage

Usage	Category
Email	
Web browsing	Accelerated Internet and HD Video
Office productivity	
Listen to music	Simplified Content Management
Online chat	Accelerated Internet and HD Video
Watching online video	Accelerated Internet and HD Video
Photo editing	Simplified Content Management
Personal finances	
Taking notes	
Online web-based games	Immersive Gaming
Social networking	Accelerated Internet and HD Video
Calendar management	
Locally installed games	Immersive Gaming
Educational apps	
Video editing	Simplified Content Management
Internet phone	Accelerated Internet and HD Video

0% 20% 40% 60% 80% 100%

New Experiences

- Accelerated Internet and HD Video
- Simplified Content Management
- Immersive Gaming

Source: IDC's 2009 Consumer PC Buyer Survey

People Prefer Visual Communications

Verbal Perception

Words are processed at only 150 words per minute

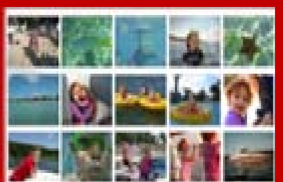
Visual Perception

Pictures and video are processed 400 to 2000 times faster

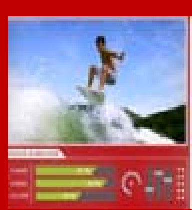
- Rich visual experiences
- Multiple content sources
- Multi-Display
- Stereo 3D

The Emerging World of New Data Rich Applications

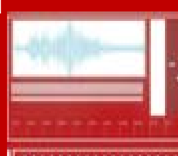
The Ultimate Visual Experience™
Fast Rich Web content, favorite HD Movies, games with realistic graphics



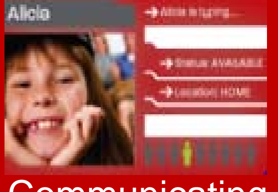
Using photos
- Viewing& Sharing
- Search, Recognition, Labeling?
- Advanced Editing



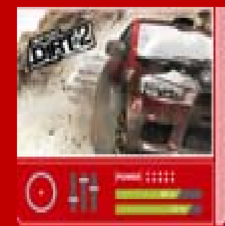
Using video
- DVD, BLU-RAY™, HD
- Search, Recognition, Labeling
- Advanced Editing & Mixing



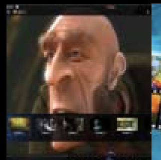
Music
- Listening and Sharing
- Editing and Mixing
- Composing and compositing



Communicating
- IM, Email, Facebook
- Video Chat, NetMeeting



Gaming
- Mainstream Games
- 3D games



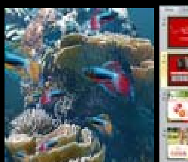
ArcSoft TotalMedia® Theatre 5

ArcSoft Media Converter® 7

ViVu Desktop Telepresence

CyberLink Media Espresso 6

Nuvixa Be Present

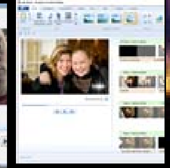


CyberLink Power Director 9

Corel Digital Studio 2010

Internet Explorer 9

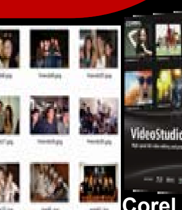
Microsoft® PowerPoint® 2010



Windows Live Essentials

Codemasters F1 2010

Viewdle Uploader



Corel VideoStudio Pro

AMD
The future is fusion

New Workload Examples: *Changing Consumer Behavior*

24 hours
of video
uploaded to YouTube
every minute

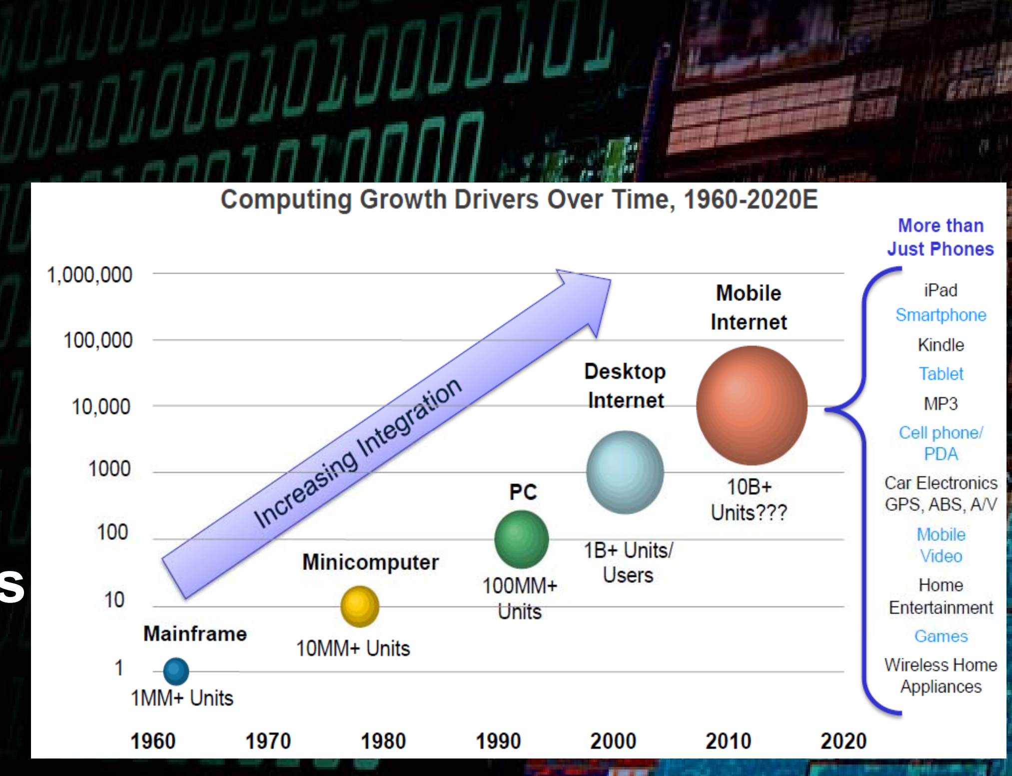
Approximately
9 billion
video files owned are
high-definition

50 million +
digital media files
added to personal content libraries
every day

1000
images
are uploaded to Facebook
every second

What Are the Implications for Computation?

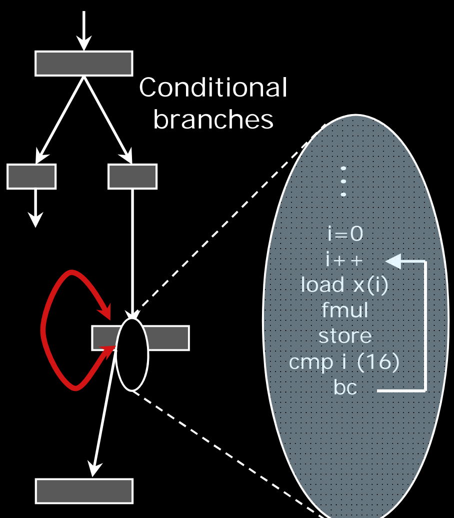
- **Insatiable demand for high bandwidth processing**
  - Visual image processing
  - Natural user interfaces
  - Massive data mining for associative searches, recognition
- **Some of these compute needs can be offloaded to servers, some must be done on the mobile device**
  - Similar compute needs and massive growth in both spaces



How must CPU architecture change to deal with these trends?

Serial Computation

Serial Code



Loops, branches and conditional evaluation

35 Years of Microprocessor Trend Data

10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^0

1975 1980 1985 1990 1995 2000 2005 2010 2015

Transistors (thousands)

Single-thread Performance (SpecINT)

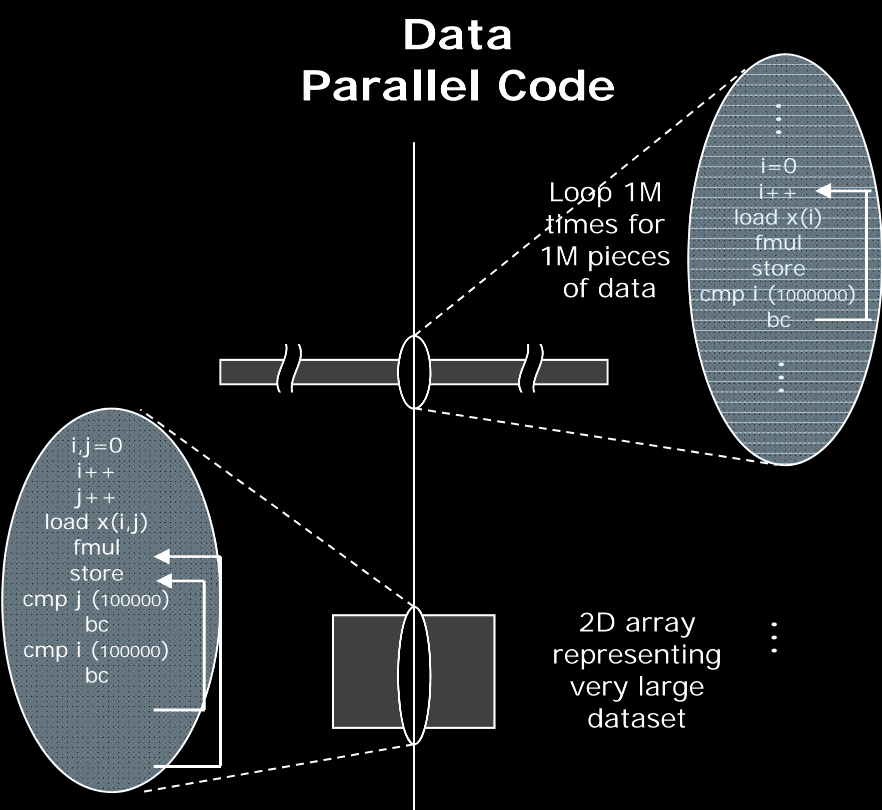
Frequency (MHz)

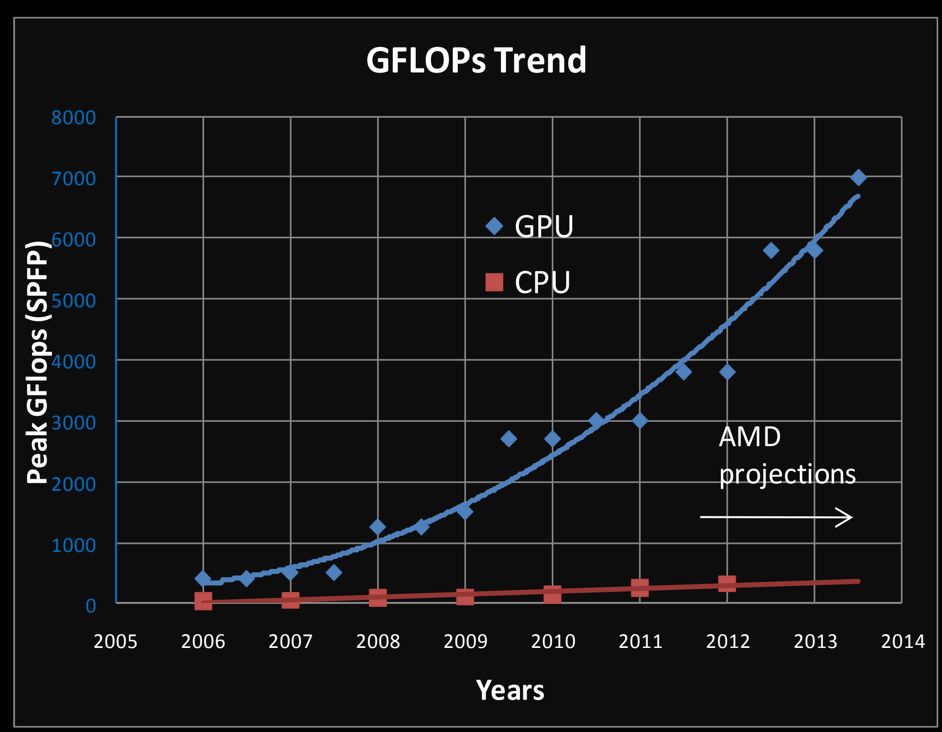
Typical Power (Watts)

Number of Cores

Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten

Parallel Computation





GPU/CPU Design Differences

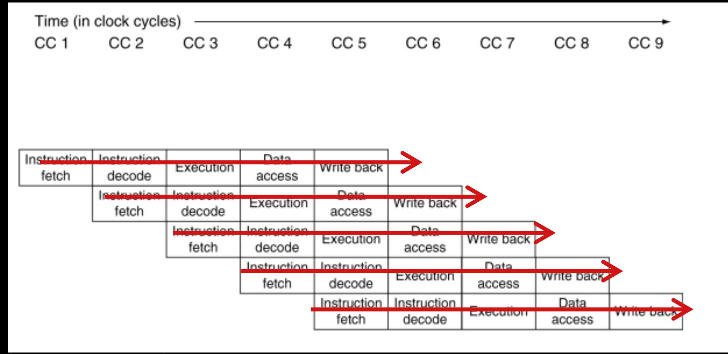
CPU (Serial compute)

Lots of instructions little data

- Out of order exec, Branch prediction
- Few hardware threads

Weak performance gains through density

Maximize speed with fast devices



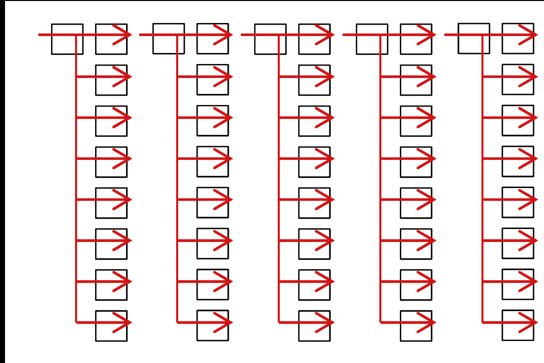
GPU (parallel compute)

Few instructions lots of data

- Single Instruction Multiple Data
- Extensive fine-threading capability

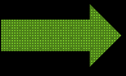
Nearly linear performance gains with density

Maximize density with cool devices



Three Eras of Processor Performance

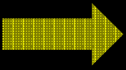
Single-Core Era



Enabled by:
- ✓ Moore's Law
- ✓ Voltage & Process Scaling
- ✓ Micro Architecture

Constrained by:
- ✘ Power
- ✘ Complexity

Multi-Core Era



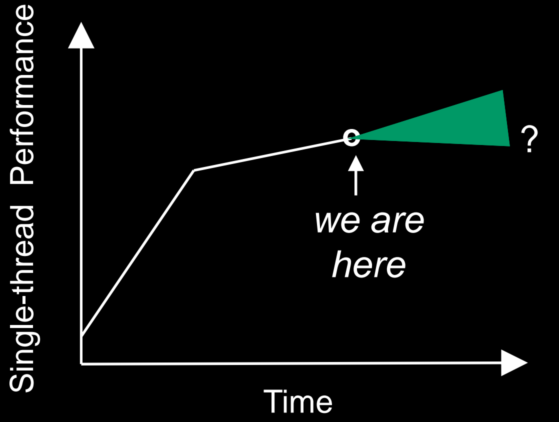
Enabled by:
- ✓ Moore's Law
- ✓ Desire for Throughput
- ✓ 20 years of SMP arch

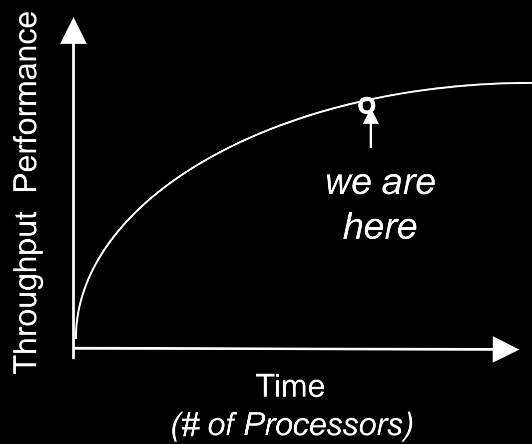
Constrained by:
- ✘ Power
- ✘ Parallel SW availability
- ✘ Scalability

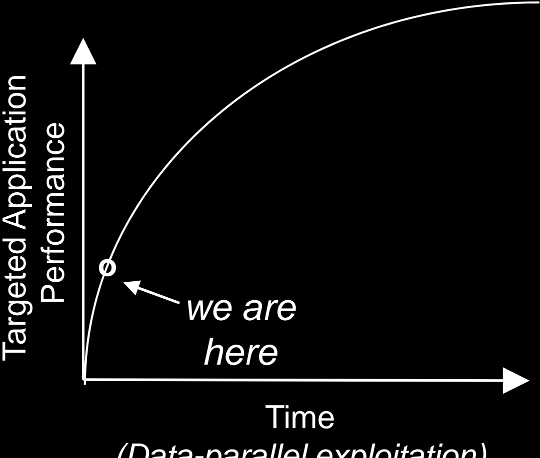
Heterogeneous Systems Era

Enabled by:
- ✓ Moore's Law
- ✓ Abundant data parallelism
- ✓ Power efficient GPUs

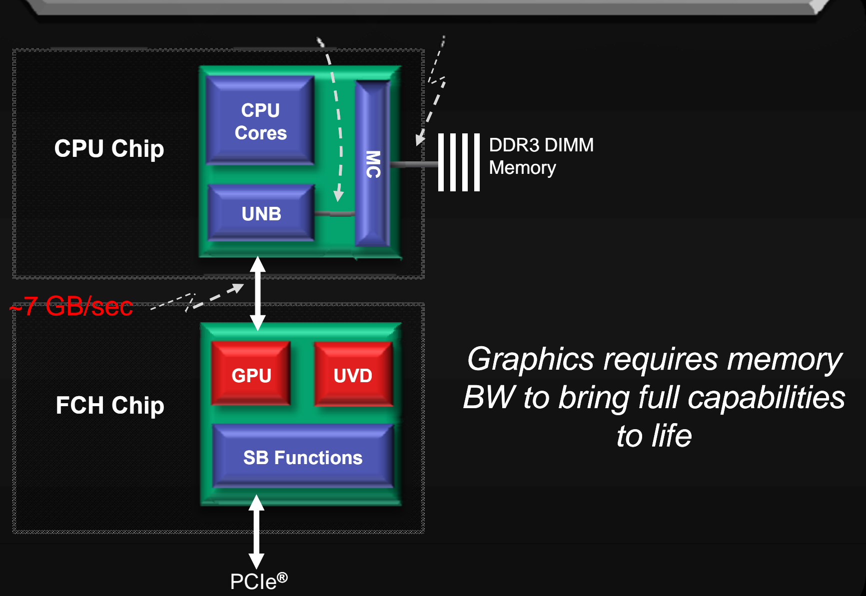
Temporarily constrained by:
- ✘ Programming models
- ✘ Communication overheads
- ✘ Workloads

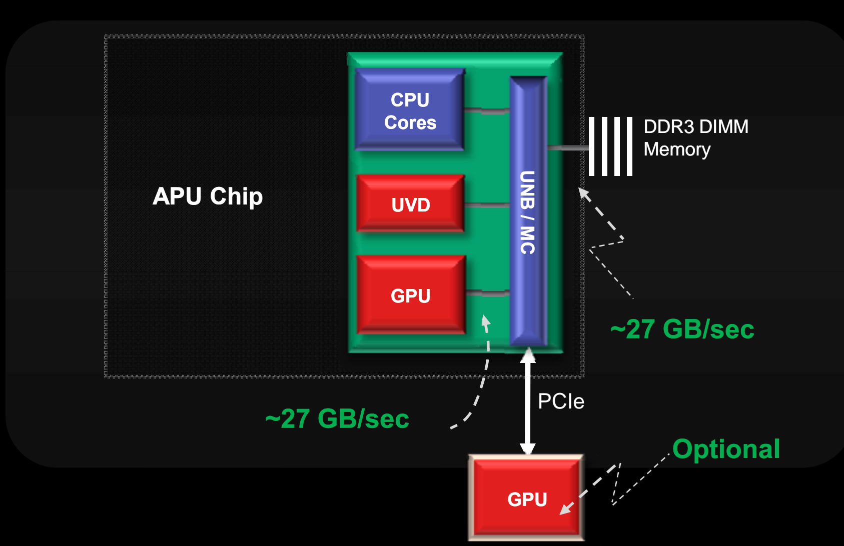






Heterogeneous Computing with an APU Architecture





Bandwidth pinch points and latency hold back the GPU capabilities

Integration Provides Improvement

- Eliminate power and latency of extra chip crossing
- 3X bandwidth between GPU and Memory!
- Same sized GPU is substantially more effective
- Power efficient, advanced technology for both CPU and GPU

The Challenges of Integration

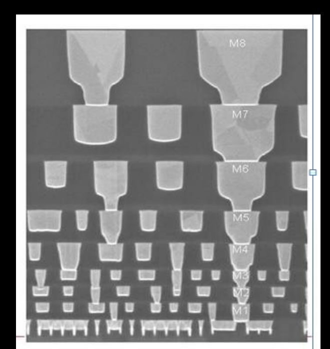
Performance

GPU

CPU

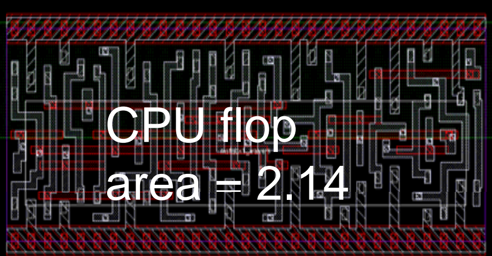
Density

Thick, fast metal
Big devices

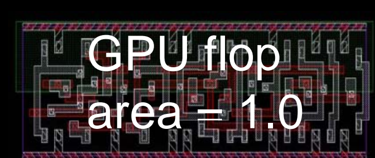


Dense, thin metal, small devices

Flop count for 4 Llano CPU cores=0.66M



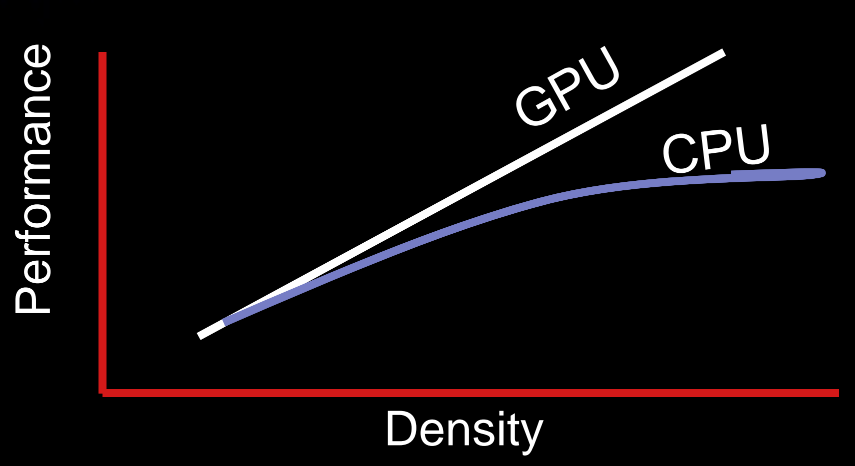
CPU flop area = 2.14

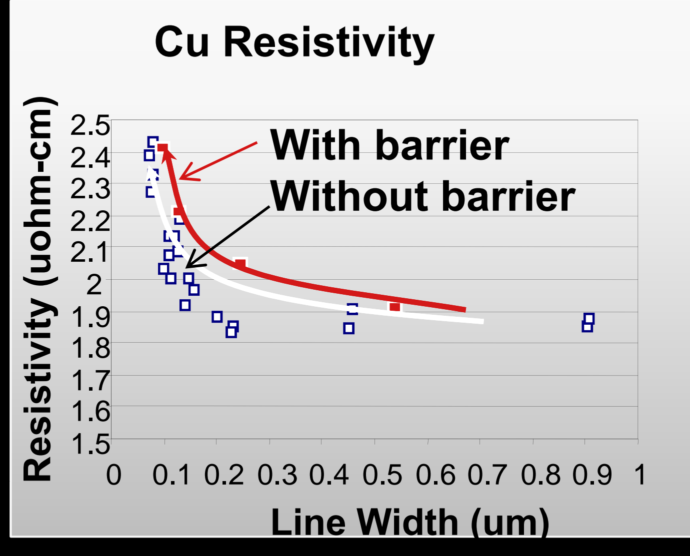


GPU flop area = 1.0

Flop count for Llano GPU =3.5M

How to Balance the Metal Stack?





With the 20nm node, even local metal will be seeing large RC increase → compromises more difficult

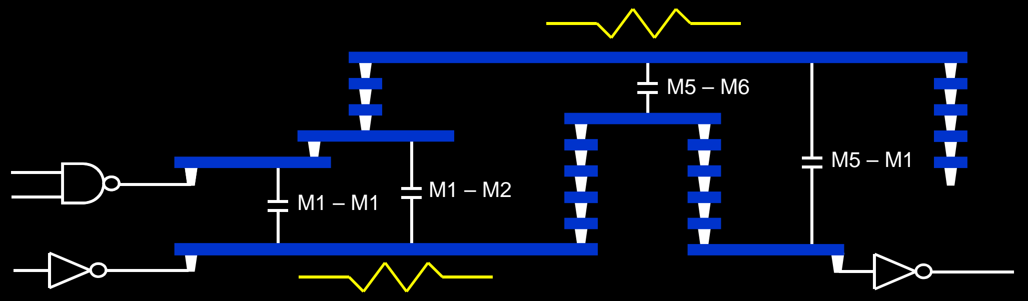
Add metal layers?

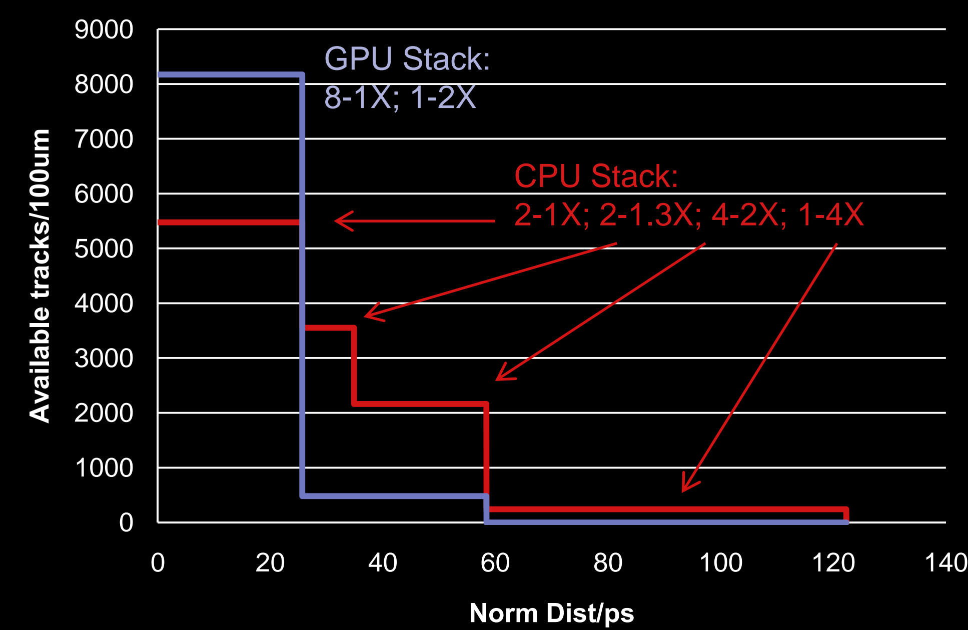
- Thin, dense layers for the GPU
- Thick, low resistance layers for the CPU
- Cost issues?
- Via resistance?

Technology improvements in BEOL are required

R vs C?

- Given the grim RC prognosis, should we be re-shaping either the aspect ratio or stack composition?
- Maybe.
- However, there are times when RC is important, but there are also many times when only C matters
- Moreover, metal stack aspect ratio is more or less maxed out, so that leaves stack composition
- Different products will emphasize different metal stacks





The growth in metal layer count

Number of CPU Metal Levels vs Technology Node

Metal Levels: 0, 2, 4, 6, 8, 10, 12, 14

Technology Node: 10, 100, 1000

Factors driving growth in Metal Layers

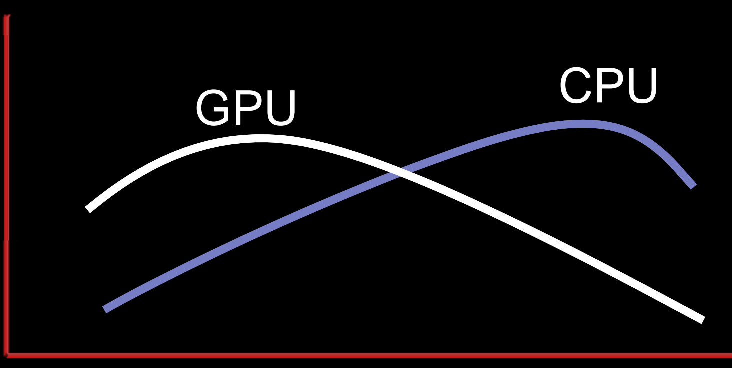
- Interconnect requirements from basic scaling
  - Transistor count N scales as S^2 (with fixed die size)
    - Total interconnect length (in lambda) scales $N^{>1}$ because of semi-global and global routes. Therefore, interconnect length (in mm) increases at a rate $<1/S$
- Non-scaling design rules
  - In order to achieve tight pitch, more restrictive design rules are imposed that significantly reduce the routeability of metal layers:
    - Unidirectional metal, increased overlap requirements, restrictive T2T and T2L rules
    - Each metal layer is "worth less" in terms of routeability: need more metal layers
- Reverse scaling
  - Long distance routes require lower RC than can be accommodated by scaled metal
  - So, move routes to thicker layers, but fewer tracks available, so pressure on layer count

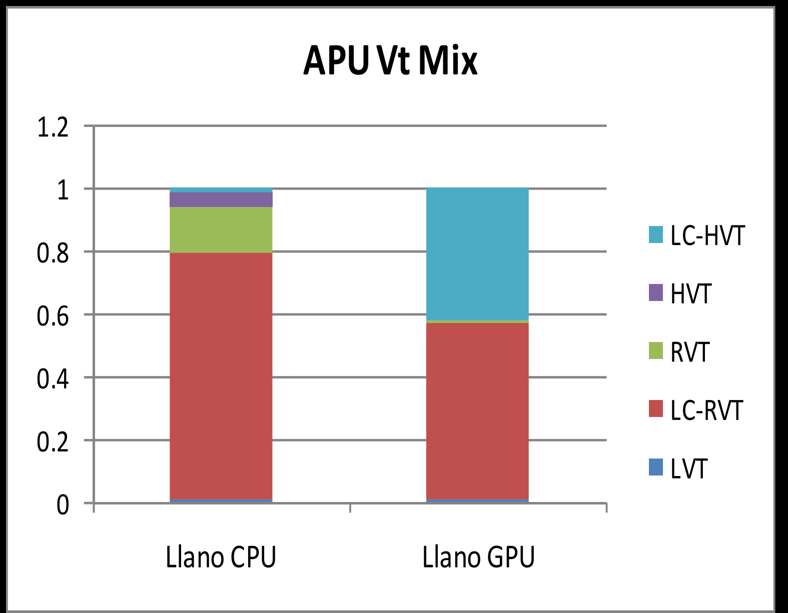
Factors driving increase in Metal Layers

- Electromigration/Power Supply Grid
  - As cross section scales with S^2, and current increases as V_{dd} drops, so current densities increase dramatically
    - Higher Via R, Metal resistances significantly degrade
    - Drives improved E-M sophistication, process techniques (alloys/barriers), denser power networks
    - Power Gating and Power Islands may drive the need for multiple supply grids
    - **More metal consumed by power supply grid**

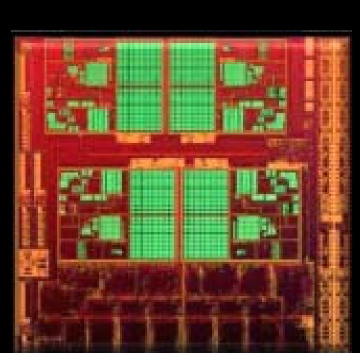
- All of the above can have the effect of increasing the number of metal layers
  - But it can be a tradeoff of Metal layers vs die size and/or route time

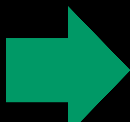
Device Optimization

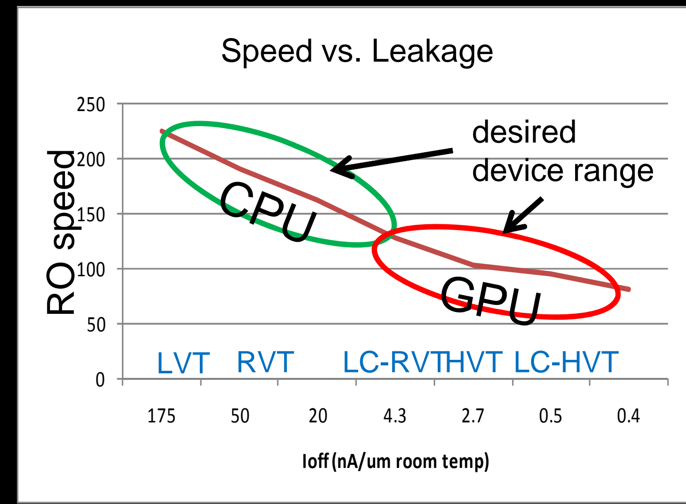




To achieve breakthrough APU performance, the Llano GPU has ~5X the flops and ~5X the device count of the CPUs







A broader device suite is required

Power Transfers

CPU Core+ cache
CPU Core+ cache
CPU Core+ cache
CPU Core+ cache
UNB/MC
DDR IO
PCIe IO
GPU
UVD

Temperature
110.0
105.0
100.0
95.0
90.0
85.0

GPU-centric data parallel workload

Voltage range is critical to enabling the efficient power transfers that make for compelling APU performance

Power Transfers

CPU Core+ cache
CPU Core+ cache
CPU Core+ cache
CPU Core+ cache
UNB/MC
DDR IO
PCIe IO
GPU
UVD

Temperature
110.0
105.0
100.0
95.0
90.0
85.0

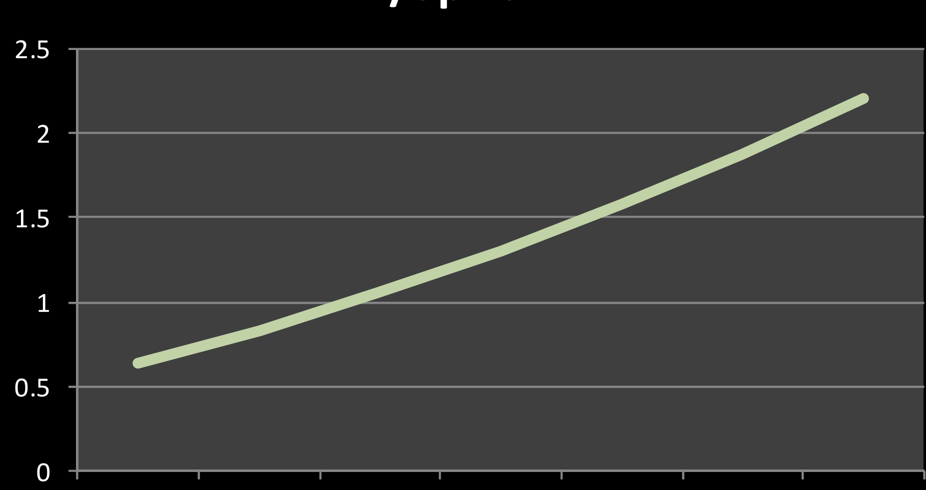
Balanced workload

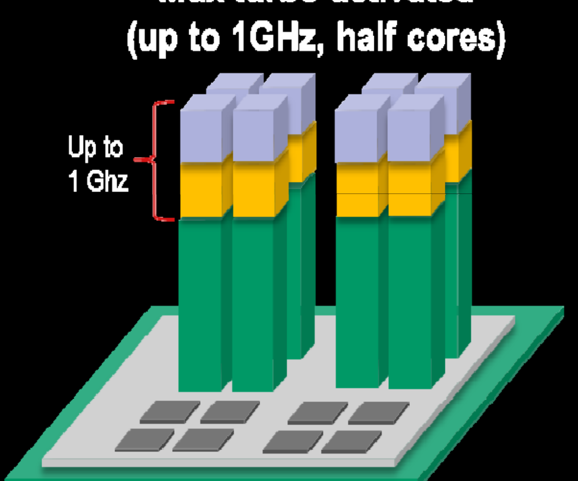
Voltage range is critical to enabling the efficient power transfers that make for compelling APU performance

Operating Voltage Range

Operating voltage requirements:

- Low voltage necessary for power efficiency
- High voltage necessary for a snappy user experience enabled by turbo mode



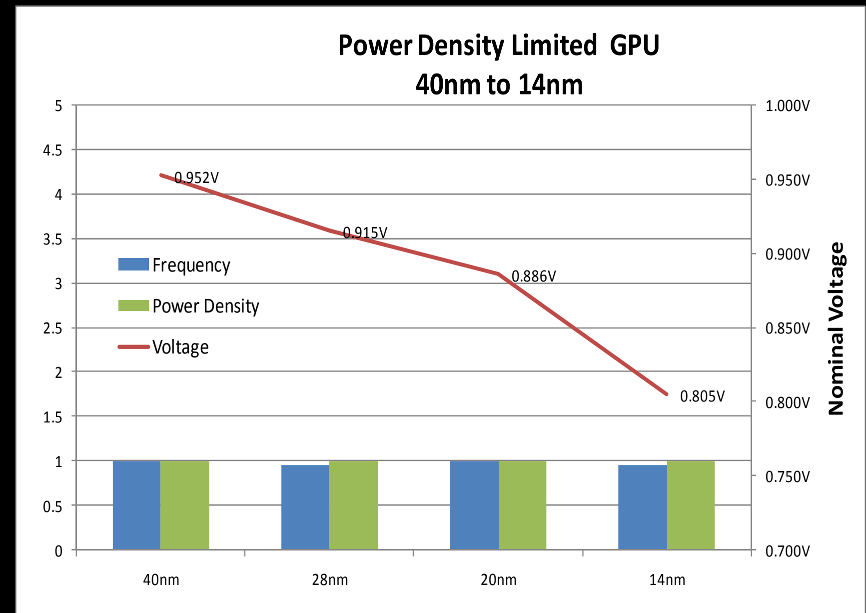


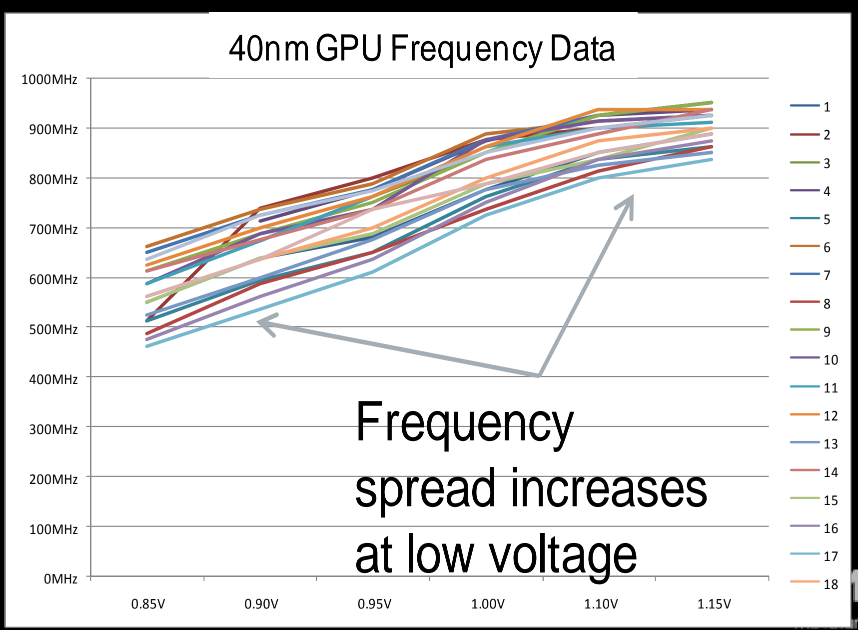
Operating Voltage Challenges

- To maintain cost effective performance growth with technology node, the GPU must:
  - Hold power density constant
  - Exploit density gains to add compute units

➔ This necessarily drives operating voltage down

- This would be good for energy efficiency except …
  - Variation impacts are much greater at low voltage





The Operating Voltage Challenge

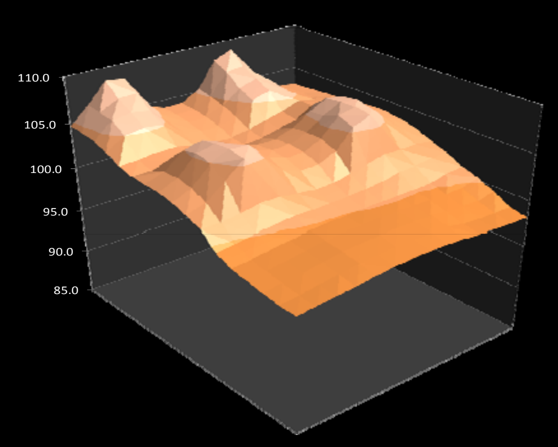
Many barriers to maintaining both high and low voltage as technology scales

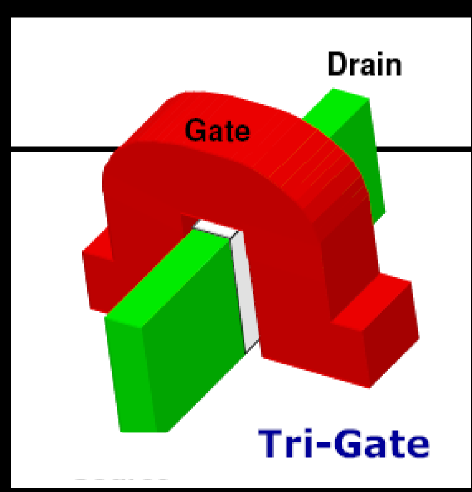
- TDDB vs. SCE control
- ULK breakdown vs. denser pitches
- Variation control

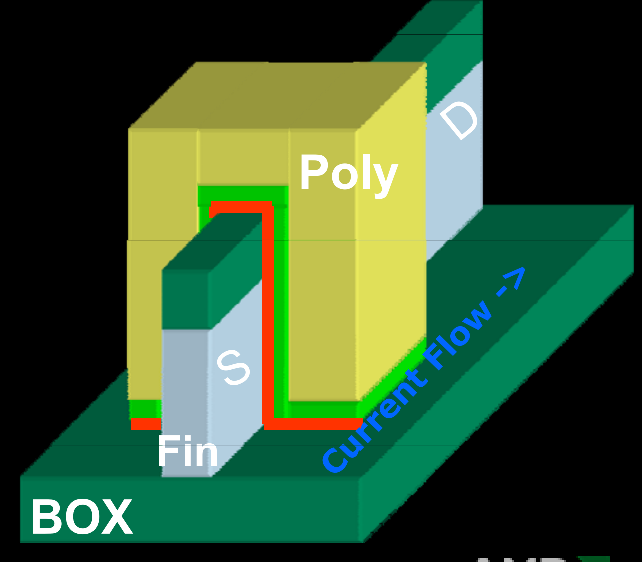
FD devices should enable maintaining the functional range for a generation or two

Will turbo modes be too compromised?

What's next?







AMD
The future is fusion

Cost issues

Lithography evolution

$$R = k_1\lambda/NA$$

λ is saturating

NA is saturating

k1 limit is about 0.25

λ

1000

100

10

ArF 193 nm immersion

ArF 193 nm

KrF 248nm

i-line 365nm

g-line 430nm

k1 reduction

NA~1.35

NA< 0.8

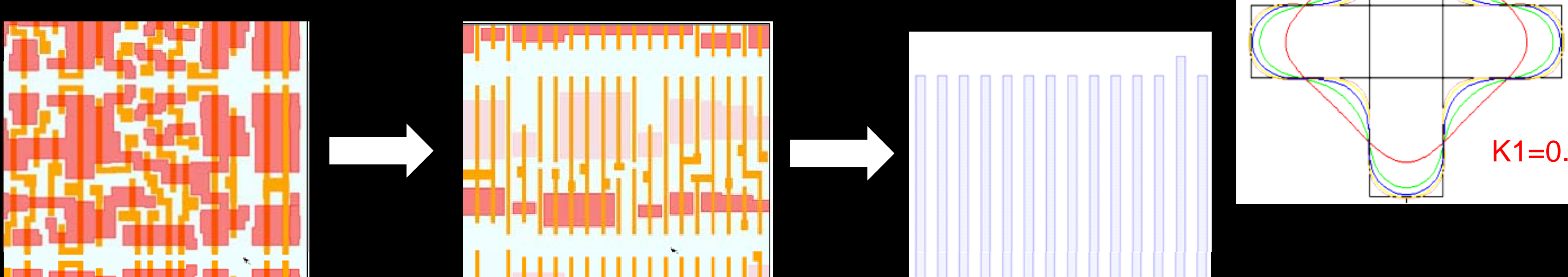
10

100

1000

Scaling implications

- $R = k_1\lambda/NA$.
  - λ stuck at 193nm for now, NA at 1.35, and k_1 limit at 0.25
  - Reducing k_1 to <0.3 has very considerable cost:
  - Much OPC and RDR needed to achieve tight pitches



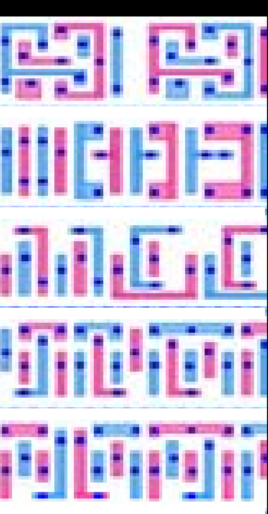
  - Net, a significant erosion of pitch-based scaling entitlement
    - Scale factors are *proprietary* … but block area scaling > pitch scaling^2!

- ***Fundamental pitch limitation for 193nm lithography is ~ 80nm***

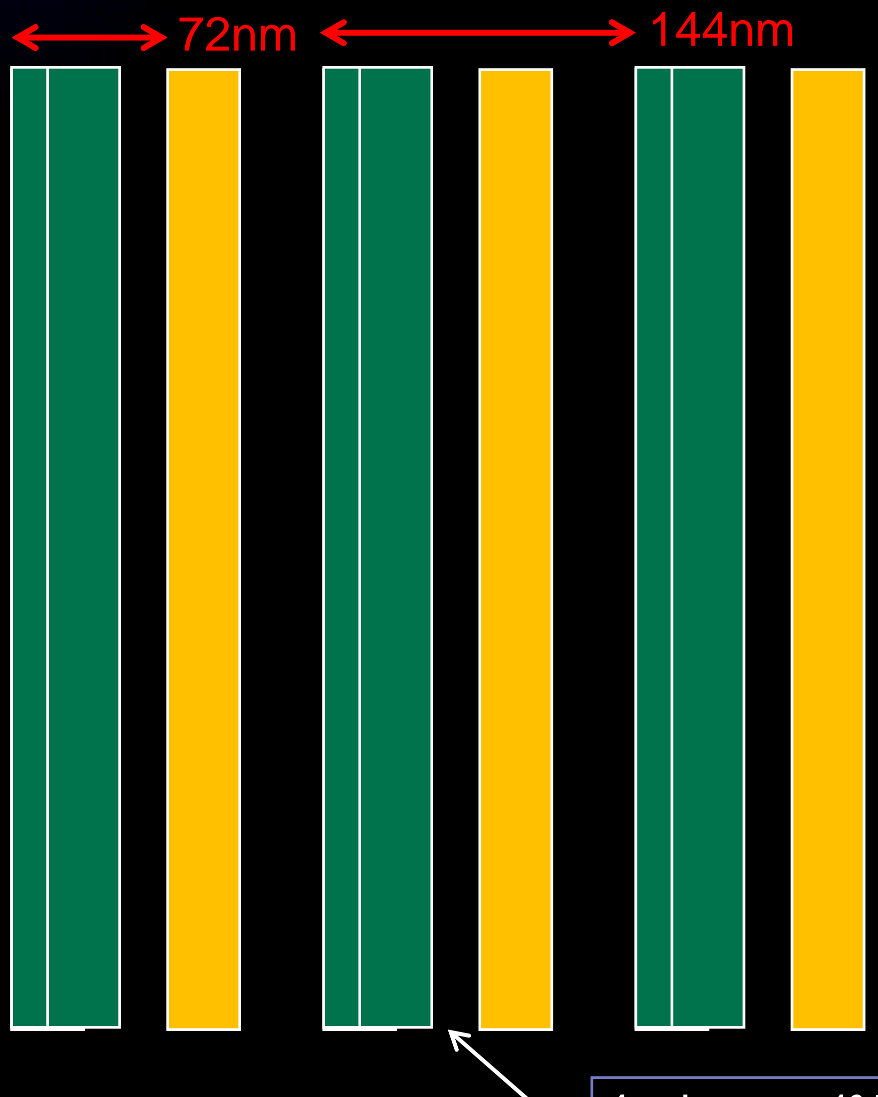
Pitch splitting

72nm 144nm

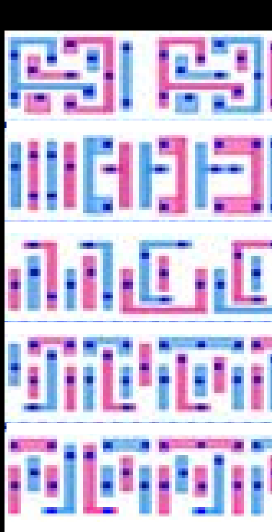
- Decomposing a layer into two effectively doubles pitch, resolving k_1 issue and allowing complex shapes
- Decomposition requires significant CAD effort to break the patterns into two printable layers



Pitch splitting



- Decomposing a layer into two effectively doubles pitch, resolving k_1 issue and allowing complex shapes



- Decomposition requires significant CAD effort to break the patterns into two printable layers

- However, now have within-layer overlay issues, and min space can be a Vmax issue, or a Cap issue

4σ min space ~ 16.5nm (PS @ 72nm) versus ~28nm (@ 80nm)

4σ max space ~ 44.8nm (PS @ 72nm) versus ~41.3nm (D@ 80nm)

→ ~Ccap variation: +85% / -30% over nominal for PS @ 72nm

→ ~Ccap variation: +25% / -15% over nominal for SE @ 80nm

Why do we care?

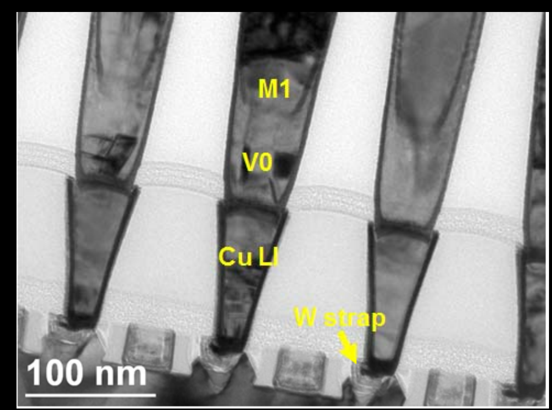
- Foundries have settled on a 28nm node with a ~4:3 M1X:Poly pitch ratio
  - Typical Design rules assuming 0.7x scaling

Design Rule	28nm	Desired 20nm
Contacted Poly Pitch	~113nm	~80nm
M1X Pitch	~90nm	~64nm

- 20nm node CPP is doable
  - but probably want >80nm for margin and gate oversize capability
- Desired 1X metal scaling to 20nm is below pitch split limit
- Can get “true” scaling and pitch split 1X metals
  - GPU’s have up to 8 1X metals
  - CPU’s have 2-5 1X metals

- ***Choice: significant cost adder for “true” scaling, or reduced cost and reduced scaling***

Other cost considerations



- MOL: Conventional contacts at <90nm CPP don’t work, and a more complex scheme is required, analogous to LI used by Intel at 32nm (+2 masks)

- BEOL Options:
  - Only scale 1X metals to ~80nm pitch, get reduced scaling but lower cost
  - Add metal layers at 80nm pitch to recover scaling; increased cost and cycle time
  - Use some combination of pitch split and non-pitch split layers to obtain greater scaling at higher cost

- Key questions to resolve:
  – Additional cost of pitch split layers
  – Additional defectivity of pitch split layers (~64 vs ~80nm pitch)
  – Whether or not to pitch split vias

Relative cost experiment

Technology Based 28nm to 22nm/20nm Die Cost and Scaling Comparison

Large Die Dominated by Wafer Cost Increase + Yield Loss (additional layers/complexity in process)

Small Die Cost Dominated by Wafer Cost Increase

38% More Logic & SRAM*

M1-M8 @ 64nm Pitch

41% More Logic & SRAM*

M1-M10 @ 80nm Pitch

Different Slopes Due to Yield Loss Over Die Size (additional layers in process)

M1-M8 @ 80nm Pitch

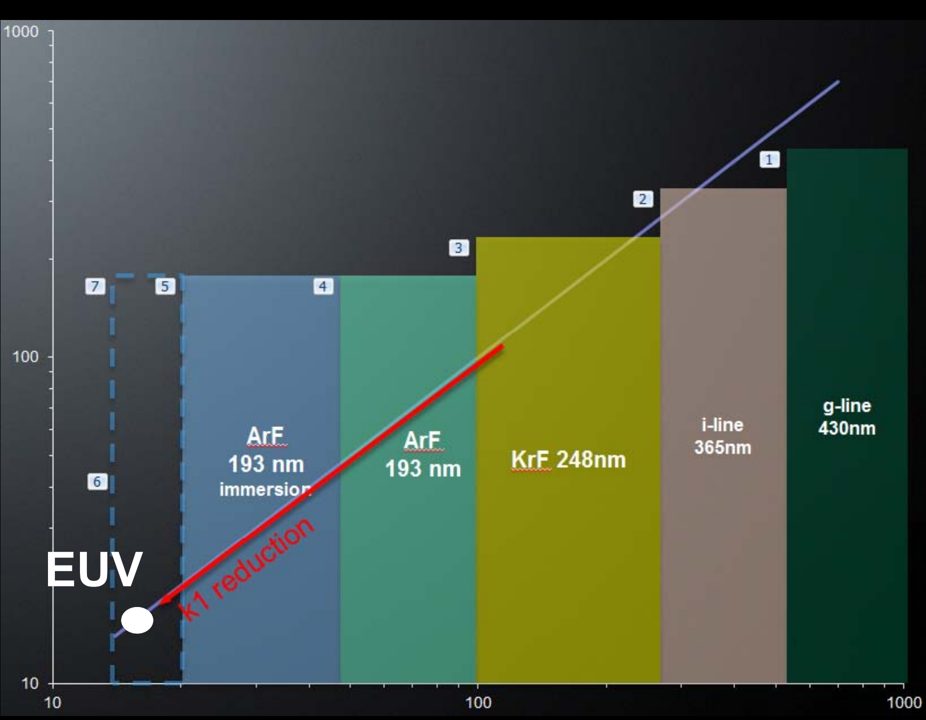
32% More Logic & SRAM*

% Die Cost Increase Normalized by Die Size

Die Size

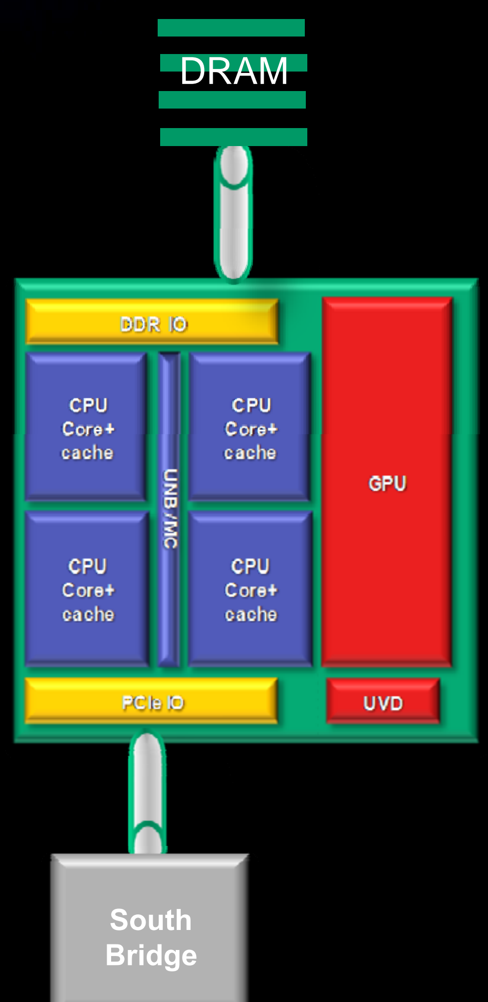
What about EUV?

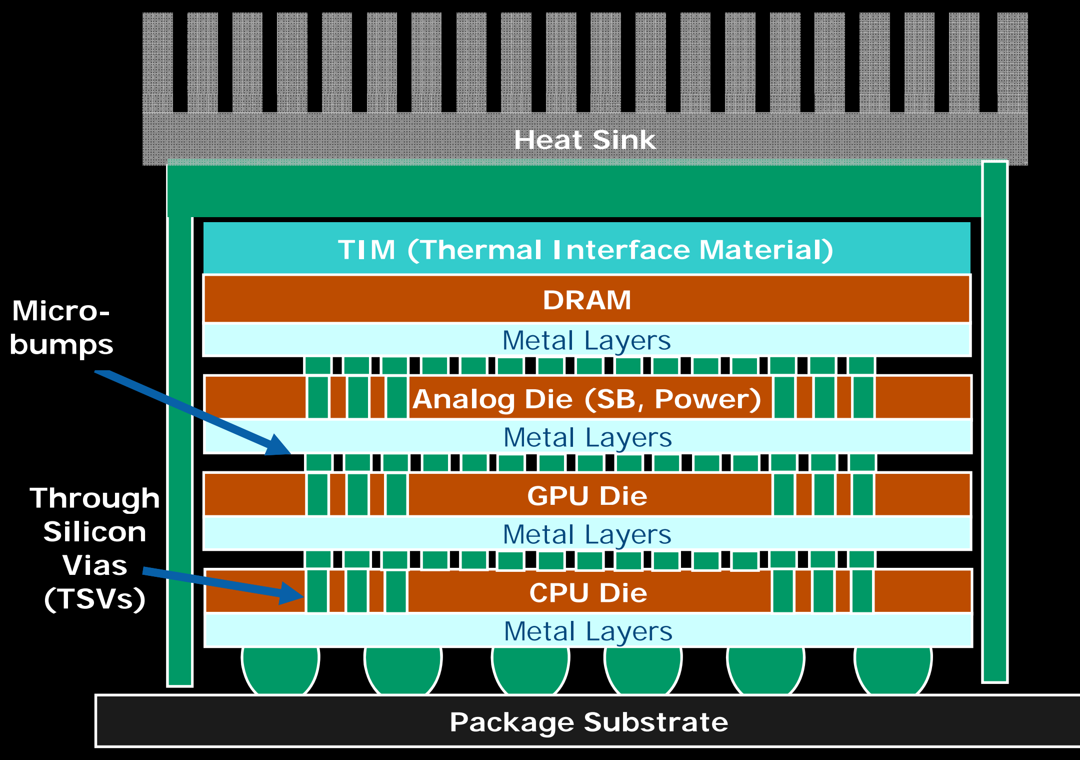
- At λ = 13.5nm, EUV should make lithography simple, and eliminate the need for pitch splitting, as well as most OPC. Right?
- Maybe:
  - Very expensive capital equipment
  - Complex, expensive reflective masks
  - Very low throughput due to illuminator output >10X below requirements
  - Very high power requirements



- These issues may be solvable, unlikely by the leading edge of 14nm
- Other forms of advanced lithography such as MEBL look attractive, but are even further behind EUV.

3D Integration to the Rescue?



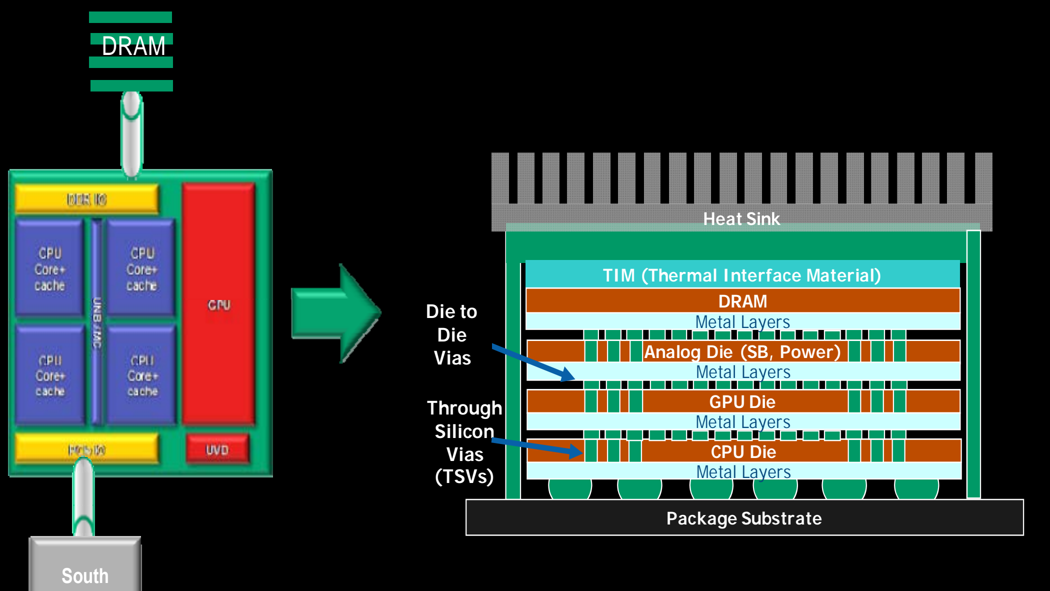


3D Integration to the Rescue?

- Stacking offers many attractive benefits
  - Higher bandwidth to local memory
  - Enables parallel and serial compute die to be in their own separate optimized technology – interconnect speed vs. density, device optimization etc.
  - Allows IO and southbridge content to remain in older, more analog-friendly technology

3D Integration Challenges

- Economical 3D stacking in high volume manufacturing presents many challenges
  - Benefits must exceed the additional costs of TSVs, and yield fallout
  - Logistics of testing and assembling die from multiple sources can be immense
  - Countless mechanical and thermal issues to solve in high volume mfg



Clearly 3D provides compelling solutions to many problems, but the barriers to entry mean heavy R&D $$ and partnerships required

Summary

- **Insatiable demand for high bandwidth computation**
  - Visual image processing
  - Natural user interfaces
  - Massive data mining for associate searches, recognition
- **Some of these compute needs can be offloaded to servers, some must be done on the mobile device**
  - Similar compute needs and massive growth in both spaces
  - Combined serial and parallel computation architectures are key in both spaces
- Huge technology challenges to meeting this opportunity
  - Interconnect scaling is hitting a wall that must be overcome
  - A broad device suite is necessary that operates efficiently at low voltage while enabling high speed for response time
  - Cost issues present a very real barrier to further scaling
  - 3D integration offers a promising long term solution