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

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

**Now: Parallel/Data-Dense**

- 16:9 @ 7 megapixels
- HD video flipcams, phones, webcams (1GB)
- 3D Internet apps and HD video online, social networking w/HD files
- 3D Blu-ray HD
- Multi-touch, facial/gesture/voice recognition + mouse & keyboard
- All day computing (8+ Hours)

*Resting battery life as measured with industry standard tests.
Focusing on the experiences that matter

Consumer PC Usage

New Experiences

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

Source: IDC's 2009 Consumer PC Buyer Survey
People Prefer Visual Communications

Words are processed at only 150 words per minute

Pictures and video are processed 400 to 2000 times faster

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

Augmenting Today's Content:
- Multi-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
CyberLink Power Director 9
Nuvixa Be Present
Corel Digital Studio 2010
Microsoft® PowerPoint® 2010
Windows Live Essentials
Codemasters F1 2010
Viewdle Uploader
Corel VideoStudio Pro

6 VLSI Technology Symposium | June 2011 | Public
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

35 Years of Microprocessor Trend Data

Transistors (thousands)
Single-thread Performance (SpecINT)
Frequency (MHz)
Typical Power (Watts)
Number of Cores

Loops, branches and conditional evaluation

Serial Code

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

Data Parallel Code

Loop 1M times for 1M pieces of data

i,j=0
i++
j++
load x(i,j)
fmul
store
cmp j (100000)
bc
cmp i (100000)
bc

2D array representing very large dataset

Peak GFLOPs (SPFP)

Years


GFLOPs Trend

GPU
CPU
AMD projections

i=0
i++
load x(i)
fmul
store
cmp i (1000000)
bc

2000 3000 4000 5000 6000 7000 8000
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

**Graphs**
- Single-thread Performance vs. Time
- Throughput Performance vs. Time
- Targeted Application Performance vs. Time

**Notes**
- The graph for Single-thread Performance shows a ? indicating an unknown future.
- The graph for Throughput Performance shows we are here.
- The graph for Targeted Application Performance shows we are here.
Heterogeneous Computing with an APU Architecture

2010 IGP-based (“Danube”) Platform

- CPU Chip
  - CPU Cores
  - UNB
- FCH Chip
  - GPU
  - UVD
  - SB Functions
- BV
- DDR3 DIMM Memory
- ~7 GB/sec

Graphics requires memory BW to bring full capabilities to life

Bandwidth pinch points and latency hold back the GPU capabilities

2011 APU-based (“Llano”) Platform

- APU Chip
  - CPU Cores
  - UVD
- UNB / MC
- GPU
- PCIe
- ~27 GB/sec

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

- Thick, fast metal
  - Big devices

- Dense, thin metal, small devices

Flop count for 4 Llano CPU cores = 0.66M

CPU flop area = 2.14

Flop count for Llano GPU = 3.5M

GPU flop area = 1.0
How to Balance the Metal Stack?

With the 20nm node, even local metal will be seeing large RC increase \( \rightarrow \) 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

Technology Node
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
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.
Voltage range is critical to enabling the efficient power transfers that make for compelling APU performance.
Power Transfers

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

[Graph showing E/op vs. V]
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?
Cost issues
Lithography evolution

\[ R = \frac{k_1 \lambda}{NA} \]

\( \lambda \) is saturating

NA is saturating

k1 limit is about 0.25

Technology Node or min Feat. size

ArF 193 nm immersion

i-line 365nm

KrF 248nm

g-line 430nm

NA~1.35

NA<0.8
Scaling implications

- \( R = \frac{k_1 \lambda}{NA} \).
  - \( \lambda \) 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**

- 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 $V_{\text{max}}$ issue, or a Cap issue.

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

- 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

<table>
<thead>
<tr>
<th>Design Rule</th>
<th>28nm</th>
<th>Desired 20nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contacted Poly Pitch</td>
<td>~113nm</td>
<td>~80nm</td>
</tr>
<tr>
<td>M1X Pitch</td>
<td>~90nm</td>
<td>~64nm</td>
</tr>
</tbody>
</table>

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

By Layer Defect Density Assumed Constant Across Technologies (potentially optimistic)

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

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

- M1-M8 @ 64nm Pitch
- M1-M10 @ 80nm Pitch

- 36% More Logic & SRAM
- 41% More Logic & SRAM
- 32% More Logic & SRAM
What about EUV?

- At $\lambda = 13.5\text{nm}$, 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?

- **DRAM**
- **TIM (Thermal Interface Material)**
- **Heat Sink**
- **Through Silicon Vias (TSVs)**
- **Micro-bumps**
- **CPU Die**
- **GPU Die**
- **Analog Die (SB, Power)**
- **Metal Layers**
- **Package Substrate**
- **South Bridge**
- **DDR IO**
- **CPU Core**
- **cache**
- **UMB/INC**
- **PCIe IO**
- **UVD**

Diagram showing the integration of different components using 3D technology.
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