Multicore, the Memory Wall, and Numerical Compression

Al Wegener, CTO
Samplify Systems
awegener@samplify.com

Samplify Systems, Inc.
Santa Clara, CA 95051
www.samplify.com
(888) LESS-BITS
(408) 249-1500
Agenda

• Motivation: Multicore & the Memory Wall
• CPU & GPU Memory Wall Examples
• Worrying Multicore Trends from Supercomputing
• Non-scalable Memory Wall Solutions
• Emerging (Scalable) Memory Wall Solutions
• Compression for Ints, Floats, and HD Frames
• Conclusions
Motivation – Multicore & the Memory Wall

• ~2004: As per-core clock rates exceeded ~3 GHz:
  ➢ single-core CPUs consumed too much power

• Since ~2004, to follow Moore’s Law:
  ➢ CPU and GPU vendors added more and more cores per socket

• 2012: Multi-core State of the Art:
  ➢ Intel Ivy Bridge: 10 CPU cores, each with SIMD (SSE/AVX)
  ➢ Nvidia Kepler: 1536 GPU cores (floating-point engines)

• Supercomputing trend (hundreds & thousands of cores):
  ➢ per-core utilization is below 10% and falling,
    due to memory and I/O limitations

• We ask:
  ➢ will multi-core CPUs & GPUs mirror HPC’s inefficiency?
  ➢ how to improve multi-core I/O that scales with Moore’s law?
The Memory Hierarchy

Latency

- Ethernet: [1 – 1500 B]
- Flash/SSD: [4 kB]
- Disk drives: [512 B/4 kB]
- L2 cache: [32/64 B]
- L1 cache: [32/64 B]
- Registers: [2 – 8 B]

Speed

- 2 GHz
- 1 GHz
- 500 MHz
- 100 MHz

©2012 Samplify Systems, Inc.
The Memory Hierarchy

Latency

msec
1 µsec
100 nsec
10 nsec
1 nsec

Speed

2 GHz
1 GHz
500 MHz
100 MHz

Ethernet [1 – 1500 B]
Flash/SSD [4 kB]
Disk drives [512 B/4 kB]

L1 cache [32/64 B]
L2 cache [64 B]

Registers [2 – 8 B]

©2012 Samplify Systems, Inc.

...simply the bits that matter®
Multicore CPU – Pin Count & FSB/QPI Bandwidth

<table>
<thead>
<tr>
<th>Year</th>
<th>Brand name</th>
<th># cores</th>
<th>Pin count</th>
<th>FSB or QPI, GB/sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>Pentium-III</td>
<td>1</td>
<td>370</td>
<td>0.5</td>
</tr>
<tr>
<td>2001</td>
<td>Celeron</td>
<td>1</td>
<td>479</td>
<td>1.1</td>
</tr>
<tr>
<td>2002</td>
<td>Xeon 1.6</td>
<td>1</td>
<td>603</td>
<td>3.2</td>
</tr>
<tr>
<td>2003</td>
<td>Xeon 3.2</td>
<td>1</td>
<td>603</td>
<td>4.3</td>
</tr>
<tr>
<td>2004</td>
<td>Xeon 3.6</td>
<td>1</td>
<td>604</td>
<td>6.4</td>
</tr>
<tr>
<td>2005</td>
<td>Xeon 7030</td>
<td>2</td>
<td>604</td>
<td>6.4</td>
</tr>
<tr>
<td>2006</td>
<td>Xeon X5470</td>
<td>4</td>
<td>771</td>
<td>10.7</td>
</tr>
<tr>
<td>2007</td>
<td>Xeon X7350</td>
<td>4</td>
<td>604</td>
<td>8.5</td>
</tr>
<tr>
<td>2008</td>
<td>Core i7-920</td>
<td>4</td>
<td>1366</td>
<td>19.2</td>
</tr>
<tr>
<td>2009</td>
<td>Core i7-950</td>
<td>4</td>
<td>1366</td>
<td>19.2</td>
</tr>
<tr>
<td>2010</td>
<td>Core i7-970</td>
<td>6</td>
<td>1366</td>
<td>25.6</td>
</tr>
<tr>
<td>2011</td>
<td>Core i7-3960X</td>
<td>6</td>
<td>2011</td>
<td>25.6</td>
</tr>
</tbody>
</table>
Multicore CPU – Bandwidth per Core

FSB (Front Side Bus)

QPI (QuickPath Interconnect)
Multicore GPU – Increasing PCIe and GDDR Bandwidth

<table>
<thead>
<tr>
<th>Year</th>
<th>GPU</th>
<th>Shader Procs (# Cores)</th>
<th>PCIe Speed</th>
<th>GDDR Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>GeForce PCX4300</td>
<td>2</td>
<td>40</td>
<td>5.33</td>
</tr>
<tr>
<td>2004</td>
<td>GeForce PCX5300</td>
<td>5</td>
<td>40</td>
<td>5.33</td>
</tr>
<tr>
<td>2005</td>
<td>GeForce 6500</td>
<td>7</td>
<td>40</td>
<td>5.33</td>
</tr>
<tr>
<td>2006</td>
<td>GeForce 7900GT</td>
<td>64</td>
<td>40</td>
<td>42.2</td>
</tr>
<tr>
<td>2007</td>
<td>GeForce 8800GT</td>
<td>112</td>
<td>80</td>
<td>57.6</td>
</tr>
<tr>
<td>2008</td>
<td>GeForce 9800GT</td>
<td>112</td>
<td>80</td>
<td>57.6</td>
</tr>
<tr>
<td>2009</td>
<td>GeForce GTX280</td>
<td>240</td>
<td>80</td>
<td>141.7</td>
</tr>
<tr>
<td>2010</td>
<td>GeForce GTX480</td>
<td>480</td>
<td>80</td>
<td>177.4</td>
</tr>
<tr>
<td>2011</td>
<td>GeForce GTX580</td>
<td>512</td>
<td>80</td>
<td>192.4</td>
</tr>
</tbody>
</table>

GPU “core” = pixel + vertex shaders, or “unified” shaders
Multicore GPU – PCIe and GDDR Bandwidth per Core
“As the number of cores increased, the processor power increased at substantially less than linear improvement and then decreased at an exponential rate.”

Sandia explained the decrease thusly, ‘The problem is the lack of memory bandwidth as well as contention between processors over the memory bus available to each processor.’

“Others Observe the Same Multicore Problem

“Multicore meet Memory Wall”

The future of computers - Part 1: Multicore and the Memory Wall
Allocating multicore to do useful work is similarly challenging to allocating cooks in a small kitchen - each will spend time twiddling his thumbs waiting his turn.

Russell Fish — EDN, November 17, 2011

"simply the bits that matter®"
Supercomputing Trends (1 of 2)

- ECCO is one of primary workload apps of NASA Pleiades Supercomputer
  - #7 Supercomputer in World (Top500.org Nov’11)
  - Intel Harpertown CPUs (Quad core @ 3.2 GHz)

- Algorithm is MIT’s General Circulation Model (MITgcm)
  - Uses Method of Lagrange Multipliers
  - 8 GB data set

- Sustained throughput:
  - <1.4% of peak performance

Supercomputing Trends (2 of 2)

**Exascale** ($10^{18}$ flops) needs:

- **16x more DDR bandwidth per node**
- **100x more disk bandwidth per node**

## Non-scalable Solutions *(are hitting a wall)*

<table>
<thead>
<tr>
<th>“Solution”</th>
<th>Examples</th>
<th>Problems</th>
</tr>
</thead>
</table>
| More I/O pins  | Add DDR ports
Add SerDes lanes | No room on package
More I/O pins → more power pins |
| Faster I/O pins| 1.4 GHz LVDS
10 & 40 Gbps SerDes      | Board layout issues
Expensive IP blocks
Power-hungry transceivers |
| More cache     | Sandy Bridge: 20 MB
L3 cache for 8 cores | Silicon area (cost) & power |
| Packaging      | Stacked die:
DDR + CPU + flash | Limited # of thru-silicon vias
Doesn’t solve socket-to-socket IO |
| Optical        | IBM: “maybe in 2018”            | Cost
Optical “compute” element?
E ↔ O is power-hungry |
Emerging, Scalable Multi-core I/O Improvements

- Hardware & Software Pre-fetch (DMA)
- Dedicated cores and/or threads for I/O
- Numerical Compression

**APAX:**

*APplication AXceleration*
Hardware & Software Pre-fetch

“For a given application, … [we] proactively load data into the caches using software [and hardware] pre-fetch instructions.”

“The placement of the pre-fetch instructions was chosen empirically, through experimentation … many system factors come into play.”

Memory Wall Take-away:
Optimizing data pre-fetch can make apps run faster
Dedicate Some Cores & Threads for I/O

Compute cores + I/O cores

cudaDMA

Memory Wall Take-away:
Dedicate some cores (or warps) for I/O

M. Dorier: “On the benefit of dedicating cores to mask I/O jitter in HPC simulations”

Bauer, Cook, and Khailany, “CudaDMA: Optimizing GPU Memory Bandwidth via Warp Specialization”
A Compression Product: VMware ESX

Figure 8. Host swapping vs. memory compression in ESX

- Lossless only
- Must achieve 2:1

Source: Understanding Memory Resource Management in VMware ESX 4.1

©2012 Samplify Systems, Inc.
Intel’s Variable Precision Floating Point

Justin Rattner, Intel CTO, Feb 2012:

“Another new circuit that is going to have supercomputer boffins salivating is a **variable precision floating point unit**.”

“This is a test chip from Intel Labs that sniffs the applications using the FP unit, seeing whether they are doing math for apps or for graphics processing, and **figures out what level of precision is necessary then only takes data in that format and only crunches it at that level of bit-ness.**

- 6-bit, 12-bit, and 24-bit precision
- Cuts power consumption “by up to 50%”
- Rattner: “there is no diminished resolution on images … at least not that the eye can see.”

Source: [http://www.theregister.co.uk/2012/02/19/intel_isscc_ntv_digital_radio/](http://www.theregister.co.uk/2012/02/19/intel_isscc_ntv_digital_radio/)
...perhaps IEEE-754 is sometimes overkill?

...simply the bits that matter®
APAX: a Numerical Compressor for I/O

Ints, floats, & frames
(from faster memory)

Floating-pt Processor

Redundancy Remover

Bit Packer

Packet Header Generator

Compressed APAX packets
(to slower memory)

~100k gates
FPGA latency
~30 clocks

User chooses:
• lossless
• fixed rate (0.05)
• fixed quality (0.5 dB)

DMA descriptor

APAX Control

Memory Wall Take-away:
Compression is easier than designing faster DDR

©2012 Samplify Systems, Inc.
On-chip Silicon Area for Arithmetic

CPUs & GPUs:
- FPU
- SIMD (SSE/AVX)

App Processors:
- Wireless Baseband
- ARM Neon SIMD
- H.264 frame buffers

FPGA DSP Blocks

All of these dedicated arithmetic IP blocks could be I/O-accelerated
Things that Make You Go *hmm*…

“Binary floating-point [FP] representations of most real numbers are inexact, and there is an inherent uncertainty in the result of most calculations involving floating-point numbers. Programmers of floating-point apps typically have the following objectives:

- Accuracy
- Reproducibility
- Performance

These objectives usually conflict! However, good programming practices and judicious use of compiler options allow you to control the tradeoffs. “


©2012 Samplify Systems, Inc.
On-chip Silicon for Cache \( (\text{is going “dark”}) \)

Dark Silicon and the End of Multicore Scaling

Proceedings of the 38th International Symposium on Computer Architecture (ISCA ‘11)

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.

At 22 nm (i.e. in 2012), 21% of the chip will be dark and at 8 nm, over 50% of the chip will not be utilized using ITRS scaling.
Data bandwidth dominates the performance and power consumption in the video encoder design. A low bandwidth and bandwidth aware motion estimation design enables smooth and better video quality as well as lower power consumption on data accesses.

IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 20, NO. 11, NOVEMBER 2010

**Figure 1:** Video processor with an integrated FMC encoder and decoder.

A New Frame Memory Compression Algorithm with DPCM and VLC in a 4×4 Block

Yongseok Jin, Yongje Lee, and Hyuk-Jae Lee

Department of Electrical Engineering and Computer Science, Inter-University Semiconductor Research Center, Seoul National University, Seoul 151-742, South Korea

©2012 Samplify Systems, Inc.

...simply the bits that matter®
**Lossless: OK. But LOSSY? Yes … 2:1 to 6:1**

### INTEGERS

<table>
<thead>
<tr>
<th>Signal Type</th>
<th>Lossless C. R.</th>
<th>Fixed rate C. R. &amp; quality metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wireless baseband (3G, LTE)</td>
<td>1.2:1 – 1.5:1</td>
<td>1.6:1 – 2.3:1 EVM, PCDE, ACLR</td>
</tr>
<tr>
<td>Computed tomography</td>
<td>1.6:1 – 2.7:1</td>
<td>3:1 – 4.5:1 Radiologists &amp; SSIM</td>
</tr>
<tr>
<td>Ultrasound (ADC)</td>
<td>1.5:1 – 2:1</td>
<td>2:1 – 3:1 Sonographers &amp; SSIM</td>
</tr>
<tr>
<td>Ultrasound (beamformer)</td>
<td>2:1 – 3:1</td>
<td>3:1 – 4:1 Sonographers &amp; SSIM</td>
</tr>
<tr>
<td>Images &amp; video</td>
<td>1.5:1 – 2:1</td>
<td>2:1 – 3:1 viewers, PSNR, SSIM</td>
</tr>
<tr>
<td>Oscilloscope (SerDes &amp; LVDS)</td>
<td>1.3:1 – 2:1</td>
<td>2:1 – 4:1 BER, rise/fall time</td>
</tr>
<tr>
<td>Radar</td>
<td>2:1 – 3:1</td>
<td>3:1 – 5:1 p_d, p_a</td>
</tr>
<tr>
<td>High-speed Imaging</td>
<td>1.4:1 – 2:1</td>
<td>2:1 – 4:1 Photographers, SSIM</td>
</tr>
<tr>
<td>Motion est. frames</td>
<td>1.6:1 – 2.5:1</td>
<td>2:1 – 6:1 Viewers, SSIM</td>
</tr>
<tr>
<td>Graphics textures</td>
<td>1.8:1 – 2.8:1</td>
<td>3:1 – 8:1 Gamers, SSIM</td>
</tr>
</tbody>
</table>

### FLOATS

#### GPU Application

<table>
<thead>
<tr>
<th>Fixed Rate Comp Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydrodynamics</td>
</tr>
<tr>
<td>Car crash simulation</td>
</tr>
<tr>
<td>Computational fluid dynamics</td>
</tr>
<tr>
<td>Financial models</td>
</tr>
<tr>
<td>Image recognition</td>
</tr>
</tbody>
</table>

©2012 Samplify Systems, Inc.

...simply the bits that matter®
APAX Compresses Textures *(better than DTX)*

- APAX achieves real-time encoding and decoding
  - DXTx only real-time decoding
- APAX allows system designer to control compression ratio vs. image quality
  - DXTx fixed at 3:1 compression
- APAX supports multiple color spaces
  - Bayer matrix, RGB 4:4:4, or YUV
  - DXTx only RGB
- APAX supports multiple pixel depths:
  - 8/10/12/14/16 bits per color plane
  - Texture compression only 8/16 bits

![Image Compression Study](image-url)

Univ Waterloo Image Suite, 8 bits per color

Perceptually Lossless Threshold

©2012 Samplify Systems, Inc.

...simply the bits that matter®
APAX Does High Dyn Range Images *(better)*

- APAX achieves real-time encoding and decoding
  - BC6 encoding *very slow*
- APAX allows system designer to trade off compression ratio and image quality
  - BC6 fixed at 6:1
  - APAX perceptually lossless at 8:1
- APAX supports multiple color depths
  - 8/10/12/16-bit color
  - BC6 only 16-bit color
- APAX supports multiple color spaces
  - BC6 only RGB
Real World Analog Sensors and “Over-casting”

“the Real World” (analog)

Transducer ($V_{out}$ or $I_{out}$)

A/D $N = 8$ to $N = 20$ bits (integers)

typecast for “convenience”

$x = \text{float} \ i$;
$y = \text{double} \ j$;

The Memory Wall

Multi-core CPU/GPU cache

DDRX

32 & 64 bits

©2012 Samplify Systems, Inc.

...simply the bits that matter®
APAX’s “Just Right” Numerical Resolution

“the Real World” (analog)

Real-world example: Pressure transducer

Range: 0 – 30,000 psi
Accuracy: ±3 psi
Resolution: 20-bit A/D → ± 0.001 psi

Transducer ($V_{out}$ or $I_{out}$)

A/D

N = 8 to N = 20 bits (integers)

Scaling the Memory Wall

$x = (\text{float}) i;$
$y = (\text{double}) j;$

Multi-core CPU/GPU

cache

DDR

32 & 64 bits

8 to 32 bits

©2012 Samplify Systems, Inc.

...simply the bits that matter®
Example: How Much Compression?

- 1000000 pressure measurements
- Sample value: min = 3.4253, max = 1655.7573
- Frequency: noise floor 28.7 dB, max 79.7 dB, dyn range 51.0 dB

©2012 Samplify Systems, Inc.
Lossy Error = Original - Decompressed

Input samples (ints or floats)

compress

decompress

decompressed signal

original signal

Error

psd = “power spectral density”

...an often-used Matlab function

FFT (psd)

Error spectrum

...simply the bits that matter®
Example: Error at 2:1 and 3:1 Compression

Range: \(\{281.66, 1071.13 \text{ psi}\}\)
Blue: Original
Red: Decompressed

(Do you see any blue?)

Error after 2:1 compression:
\(\{-0.164, +0.189\}, \text{ or } \{-0.02\%, +0.02\%\}\)

...vs. accuracy of ±3 psi?

Error after 3:1 compression:
\(\{-1.099, +1.113\}, \text{ or } \{-0.13\%, +0.14\%\}\)

...vs. accuracy of ±3 psi?

...simply the bits that matter®
Example: Spectra at 2:1, 3:1, and 4:1

- Orig, decomp, and error spectrum at 2:1, 3:1, and 4:1

- Dynamic range (dB): 51.0 dB

- APAX error spectra are below the noise floor:
  - at 2:1, error is ~23 dB under
  - at 3:1, error is ~22 dB under
  - at 4:1, error is ~19 dB under

- ...about 9 integer bits

- ...simply the bits that matter®
**APAX Scales** With the Memory Wall

CPU I/O: 12.8 GB/sec
- ints, floats, frames
- 240 pins @ 1.6 GHz
- 12.8 GB/sec

CPU I/O: 25.6 GB/sec
- ints, floats, frames
- 240 @ 1.6 GHz
- 2:1 compress

CPU I/O: 38.4 GB/sec
- ints, floats, frames
- 240 @ 1.6 GHz
- 3:1 compress

Scalable Solution:
- Trade gates for pins
- PHY rate stays the same
- Effective throughput increases
- “Gates for pins” knob is controlled in **software**

©2012 Samplify Systems, Inc.

...simply the bits that matter®
Samplify Partnerships

2009
- Computed Tomography
  - Prism CT
  - Exclusive licensee

2010
- Wireless
  - Prism IQ
  - Strategic investor
  - Acquired wireless infrastructure patent portfolio

2011
- High Performance Computing
  - APAX
  - Strategic investor
  - Exclusive licensee for oil & gas exploration

2011
- Imaging
  - Prism 3
  - Strategic investor

©2012 Samplify Systems, Inc.

...simply the bits that matter®
### APAX Software Products

60 to 400 Msamp/sec per x86 core

<table>
<thead>
<tr>
<th>Software Product</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>APAX library (.dll, .so)</td>
<td>APAX_profile()</td>
</tr>
<tr>
<td></td>
<td>APAX_compress()</td>
</tr>
<tr>
<td></td>
<td>APAX_decomp()</td>
</tr>
<tr>
<td></td>
<td>APAX_memcpy()</td>
</tr>
<tr>
<td>APAX MPI extensions</td>
<td>MPI_SEND_APAX_C()</td>
</tr>
<tr>
<td></td>
<td>MPI_RECV_APAX_D()</td>
</tr>
<tr>
<td>APAX file system extensions</td>
<td>fwrite_apax_c()</td>
</tr>
<tr>
<td></td>
<td>fread_apax_d()</td>
</tr>
<tr>
<td>APAX for CUDA &amp; OpenCL</td>
<td>cudaDMAcustom()</td>
</tr>
</tbody>
</table>
APAX IP Product

APAX IP & APAX SW share the same Application Programming Interface (API)

typedef struct __APAX_DMA_TRANS {
    APAX_SRC;
    APAX_DST;
    APAX_N_PKTS;
    APAX_MODE;
    APAX_PARAM;
    ...
} APAX_DMA_TRANS;

©2012 Samplify Systems, Inc.
Conclusions

• Multi-core CPUs and GPUs are here to stay
• Multi-core memory & I/O constraints are real, and worsening
• We’re out of pins, and 10 GHz SerDes is hard & runs hot
• Beware the gathering storm: “dark” silicon (cache memory)…

APAX Application Acceleration:

• Today’s dedicated “numerical silicon:” FPUs, SIMD, & DSP
• Tomorrow’s “numerical silicon” memory & I/O accelerator: APAX
• APAX IP:
  • Tiny chip area (~0.06 mm² in 32 nm)
  • Low latency (compared to DDRx RAS/CAS)
  • Customers control numerical I/O speed-up via a register
  • 2x to 6x application speed-up with equivalent results
  • “Numerical” killer app: accelerating motion compensation/estim.
For more information...

Al Wegener
Samplify Systems
Santa Clara, CA
awegener@samplify.com
408-249-1500

Samplify Systems, Inc: Intellectual Property
Solutions for Data Acquisition and Data Processing Bottlenecks
Samplify Systems, Inc. is a provider of intellectual property solutions which solve data acquisition and data processing bottlenecks. Samplify's Prem Compression addresses data acquisition bottlenecks in industrial, scientific, medical markets, while Samplify's APAX technology accelerates software applications which are performance-limited by IO, storage, or memory bandwidth bottlenecks in the computing, consumer, and mobile device markets.

Samplify is a private company based in Silicon Valley with backing from leading venture capitalists, Charles River Ventures and Formative Ventures and strategic partners Schlumberger, IDT, and Cosmo.

Samplify's APAX Technology Featured on EngineeringTV
From DesignWest 2012, Bill Wong interviews Samplify's Founder and CTO, Al Wegener, on the company's APAX technology for High Performance Computing, Big Data, and consumer electronics applications.