Luca Benini - Chief Architect, DEIS Università di Bologna & STMicroelectronics

Scalable Many-core Acceleration for Image Understanding - is CPU+GPU the answer?
IN-Stat: Growth of mobile App Procs in 2011 exceeded 43% and is forecast to grow at a 22% CAGR through 2016… and IOT is next!
STM “Platform 2012” → STHORM

**GOPS/mm² – GOPS/W**

<table>
<thead>
<tr>
<th>1</th>
<th>3</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>General-purpose Computing</td>
<td>Throughput Computing</td>
<td>SW Mixed HW</td>
</tr>
<tr>
<td>CPU</td>
<td>GPGPU</td>
<td></td>
</tr>
</tbody>
</table>

> 100

Closing The Accelerator Efficiency Gap
P2012 in a nutshell…

P2012 Fabric

Fabric Controller

Cluster

Cluster

Cluster

Cluster

3D-stackable SW accelerator

Customization design flow

 HW Optimized IPs

STI life:augmented
A Killer Application (domain) for P2012

Today EVA is:
1. FPGA (e.g. Xilinx) → high cost and high power
2. Specialized SIMD (e.g. Mobileye) → closed SW
3. DSP/CPU → limited Perf/W/mm²
4. CPU+GPU → upcoming, but not solid yet

Embedded Visual intelligence

The next killer app: Machines that see (J. Bier)
P2012 SoC in 28nm

- 4 Clusters, 69 processors
- 80 GOPS
- 1MB L2 mem
- 2D flip chip or 3D stacked
- 600 MHz typ
- < 2 W
- 3.7 mm² per cluster

Energy efficiency 40 GOPS/W $\rightarrow$ 0.04 GOPS/mW

Samples back from fab 6/12
Designing the P2012 SoC

Communication challenge
Interconnect Hierarchy

- **Cluster Interconnect → Latency**
- **Off-chip interconnect → Bandwidth**
- **Global Interconnect → Decoupling**

**L3 (DRAM)**
P2012 Fabric Memory Map

- Up to 32 clusters ➔ 544 processors
- 256 Mbyte ENCore Memory ➔ up to 4 Mbyte/cluster; today 256 KB/cluster
- GAS-NUMA architecture
- No caches!
Cluster interconnect
The cluster

- **Cluster Control**: ENCore Boot, HWPE control, etc...
- **DMA Control**: L3 ⇔ L1, L1 ⇔ L1
- **Configurable**: (N, EFUs, banking factor, ...) “Computing Farm”

**Debug**: Multicore Debug
P2012 Cluster Main Features

- HW Support for synchronization:
  - Fast barrier (within a cluster only) in ~4 Cycles for 16 processors
  - Flexible barrier ~20 cycles for 16 processors

- High level of customization though:
  - The number of STxP70 processing elements
  - The STxP70 extensions (ISA customization)
  - Memory sizes + Banking factor of the shared memory
ENCore16

Features:
- 2/3 cycles latency → 0/1 pipe stalls
- Non-blocking
- Parametric soft IP
Shared Memory

Processors

Routing Tree

Arbitration Tree

Memory Banks

P1  P2  P3  P4

B1  B2  B3  B4  B5  B6  B7  B8
P2012 Cluster Overview

Shared Memory Conflicts

Typical case
Inter-cluster interconnect
Each cluster has its own operating point tunable to the right energy budget

Inter cluster communication must support cluster decoupling
Decoupling → GALS

- Several synchronous clock domains (Synchronous Blocks) communicate through asynchronous channels
- Each SB can be clocked independently for local needs

**Pros**
- Reduced average frequency
- Reduced global power
- Globally skew tolerant
- No global clock in GALS: easier local clock design

**Cons**
- Global communication overhead from handshake protocol
- Total area, total wire length, and power penalty from protocol
- Local clock overhead
Fully asynchronous network on chip

- Standard-cell based hierarchical implementation with top-level pseudo-synchronous constraints
- Fully asynchronous routers & links: clockless high performance & low-power
- Network on Chip: Flexible packet-switched interconnect
- GALS SoC: Independent Time & power domains suitable for DVFS
- Low latency / low area GALS interface
P2012 SoC GANoC

Topology is tailored to architecture

System plug: asymmetric bandwidth

SoC ports

Sideband signaling: events, debug
System interconnect
SoC integration: Logical view

![Diagram of SoC integration](image-url)
3D: the die to die challenge

You need a Phy and at 500MHz to 1GHz this is tough, time consuming and expensive.

Alternative: Asynchronous Link (QDI implementation)

Pro:
- No more Phy 😊

Cons:
- Async logic on both sides
- Challenge:
  - Micro pipe stages need to be as close as they can be from micro bumps on both sides
Flexible SoC chiplet: 2D & 3D IO
Programming the P2012 Fabric

SW Ecosystem & Efficiency challenge
OpenCL Standard

Task Parallelism (run-to-completion)

Data Parallelism (with some-synchronisation)

OpenCL

ARM Cortex-A15

P2012

Rogue

More parallelism

More programmability

Multi-core CPU

Many-core

GPU

29
Independent Clusters + Independent processor IF
- Work-item divergence is not an issue
- P2012 supports more complex OpenCL task graph than GPUs. Both **task-level** and **data-level (ND-Range)** are possible

P2012 cores are not HW-multithreaded
- P2012 OCL runtime does not accept more than 16 work-items per work group when creating an ND-Range.
- But you can chose which work to do (e.g. using “case”) in each WI
The best way to hide memory transfer latencies when programming for P2012 is to overlap computation with DMA transfers. This technique is based on software pipelining and double buffering.
Memory Mapping and Data Movements

L3
>200 cycles

Global Memory (buffers)

Scalar/Vector load/store
async_work_group_copy
async_work_group_2d_copy
async_work_item_[2d]_copy

L1
shared 256KB

Local Memory (shared)
Private Memory (kernel stacks)
Constant Memory

Cluster

The compiler can compute accurately OpenCL-C kernel stack size!
P2012 OpenCL Programming Summary

1. Fill processors/clusters with processing
   - Fully with data parallelism when available
   - Task parallelism otherwise (mid-coarse grain)

2. Optimize the data locality
   - Use local & private as user managed cache
   - Minimize global ↔ local/private

3. Parallelize memory transfers & computation
   - Use asynchronous copies (DMA)
Performance Assessment
OpenCV on P2012

OCV functions accelerated with P2012 (transparently for the App programmer) → Standard domain specific APIs

FAST: key point detection
Visual Analytic Benchmarks

Running Time (ms)

- P2012-1 Cluster 600 MHz
- ARM CA9 Dual Core, 1GHz, Neon + FPU

Speed Up. P12-1Cluster vs ARM CA9 1Ghz

ARM Cortex A9 Dual Core

X 6 Computing Density
Scalability (PKLT)

Running time [ms]

1 cluster: varying ND range

2 clusters

Speedup vs ARM CA9

107

P2012 - 32cores
P2012 - 16cores
P2012 - 8cores
P2012 - 4 cores

ARM CA9 - 1GHz

50
27
16
9

0,0
2,0
4,0
6,0
8,0
10,0
12,0
14,0

0
20
40
60
80
100
120

P2012 - 32cores
P2012 - 16cores
P2012 - 8cores
P2012 - 4 cores

2,1
4,0
6,7
11,9

0,0
2,0
4,0
6,0
8,0
10,0
12,0
14,0

P2012 - 32cores
P2012 - 16cores
P2012 - 8cores
P2012 - 4 cores

2 clusters
Resolution: 640 x 480 pixels
System: Quadro NVS3100M, NVIDIA GTX280, STORM (P2012)
From prototype to product

Go to market challenge
P2012 in GOPS/mm$^2$

1. General-purpose Computing
   - CPU

2. Throughput Computing
   - GP-GPU

6. P2012 Space
   - SW
   - HW/SW
   - HW
   - CO
   - HC

> 100

HW IP
Opportunities

- Natural selection
  - Develop 10 apps
  - Succeed with 1 or 2
  - Go to market at premium on success

- Low-cost, power, high-volume
  - Quick derivative
  - Some flexibility to adapt to market and standard changes

```
<table>
<thead>
<tr>
<th>PE</th>
<th>PE</th>
<th>...</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>...</td>
<td>n</td>
</tr>
</tbody>
</table>

Shared L1 MEM

5x~10x

```

```
<table>
<thead>
<tr>
<th>PE</th>
<th>PE</th>
<th>...</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>...</td>
<td>n</td>
</tr>
</tbody>
</table>

Shared L1 MEM

```

```
<table>
<thead>
<tr>
<th>PE</th>
<th>PE</th>
<th>...</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>...</td>
<td>n</td>
</tr>
</tbody>
</table>

Shared L1 M

1.2x~3x

1x

---

STI

life.augmented
Summing Up

Heterogeneous Computing

Homogeneous Computing

- IP.s for SoCs
  - Video Codecs
  - Imaging
  - Base Band
  - IQI
  - ...

Standalone SoC
  - Eco System
  - Analytics
  - Fragmented Mkts

- GP-SMP
- GP-GPU
- HW-Pipeline
- P2012-SW
- P2012-HWSW

X Gops Small LowPower