Luca Benini
ETH Zürich & Università di Bologna

Heterogeneous Interconnects for Heterogeneous Computing
The Twilight of Moore’s Law: Economics

Market volume wall: only the largest volume products will be manufactured with the most advanced technology
The Twilight of Moore’s Law: Power

**Thermal wall:** transistor count still increases exponentially but we can no longer power the entire chip (voltages, cooling do not scale)

[Watanabe et al., ISCA’10]

[Hardavellas11]
The twilight of Moore’s Law: Communication

Memory wall: larger datasets and limited bandwidth at high power cost for accessing external memory
Market volume & power wall

Embedded Intelligence + Agile customization
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!
STMicroelectronics’ Platform 2012

GOPS/mm² – GOPS/W

General-purpose Computing

CPU

Throughput Computing

GPGPU

Mixed

SW

HW

1GOPS/mW

Platform 2012

HW IP

Closing The Accelerator Efficiency Gap with Agile Customization
Scalable, Agile Accelerator

P2012 Fabric

Homogeneous accelerator

Customization design flow
Platform 2012 → **STHORM™**

**Xilinx Zynq 7020 FPGA**
- 28 nm
- CA9, dual core, 800 MHz
- Many digital interfaces
- 1.3 Mgates equivalent FPGA

**P2012 SoC in 28nm**
- 23x23 flipchip, 600 MHz
- 2W TDP, 3.7 mm² per cluster
- 40GOPS/W → 0.04GOPS/mW
- Open source runtime+drivers
- Programmable in OpenCL

**First product EVSOC in 28nm SOI – Q4’13**
P2021/STHORM Architecture

Variations on a theme: *Heterogeneity*
A tightly-coupled computing domain, sharing
  • Local controller
  • Memory
  • Fabric interface
Strong physical & electrical decoupling (clock, power)

A loosely-coupled computing domain, sharing
  • Asynchronous NoC
  • Host interface –with configuration controller
  • System bridges to L3 memory and other SoC mega-IPs
P2012 Cluster Overview

P2012 Cluster Architecture

- N x STxP70 Cores
- 2xN-banked Shared Data Memory
- N-to-2M Logarithmic interconnect (memory)
- Peripheral Logarithmic interconnect
- Runtime accelerator (HWS)
- Timers
- Cluster interfaces (I/O)
P2012 Cluster Overview

P2012 Cluster Architecture

- 1 STxP70-based Cluster processor
- 16KB P$ & TCDM
- K DMA channels
- 1 NMI, 31 IT Interrupt controller
- CC peripheral (boot, ...)
- Clock, variability, power controller (CVP)
- Cluster Controller Interconnect

Global Interconnect Interface

Cluster Controller Interconnect (CC)

(ENCORE <N>)
P2012 Cluster Overview

P2012 Cluster Architecture

- Multi-core Sub-system
- Cluster Controller
- Debug and Test Unit (DTU)

• Provides controllability and observability to the application developer
• Breakpoint propagation inside the cluster and across the fabric
P2012 Cluster Overview

P2012 Cluster Architecture

Custom HW Processing Elements

- P x HW Processing Elements
- Stream Flow Local Interconnect (LIC)
- HWPE to/from LIC interfaces (HWPE_WPR)
- CC to/from LIC interface (SIF).

Multi-core Sub-system (ENCore <N>)

Debug and Test Unit (DTU)
## Silicon Efficiency @ Core Level

### Instruction Level Parallelism

- $X = Y + Z \quad || \quad T = U \times V$

### Data Level Parallelism

- $Vz[1..8] = Vx[1..8] + Vy[1..8]$

### Instruction Optimization

- $\langle x_1, \ldots, x_n \rangle = f(\ldots)$
XP70: core-level efficiency

- STXp70 is an ASIP “on steroids”
  - Dual issue
  - Instruction specialization
  - \textit{VECx} vector extension / \textit{FPX} floating point extension

STXp70 is an ASIP “on steroids”
- Dual issue
- Instruction specialization
- \textit{VECx} vector extension / \textit{FPX} floating point extension

Instruction-level acceleration
Cluster-level efficiency

- Tailored processor configuration
  - Dual issue + FPX

- 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
  - Memory sizes + Banking factor of the shared memory

<table>
<thead>
<tr>
<th>Feature</th>
<th>DMIPS S/MHz</th>
<th>CoreMark /MHz</th>
<th>FMax C65-LP</th>
<th>Context Switch</th>
<th>Area (C65-LP)</th>
<th>Power (C65-LP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>STxP70 V4C</td>
<td>1.4</td>
<td>2.70</td>
<td>480 MHz</td>
<td>14 cycles</td>
<td>0.15 mm²</td>
<td>60 μW/MHz</td>
</tr>
<tr>
<td>Cortex-R4</td>
<td>1.6</td>
<td>2.29*</td>
<td>500 MHz</td>
<td>30 cycles</td>
<td>0.45 mm²</td>
<td>120 μW/MHz</td>
</tr>
<tr>
<td>Cortex-M3</td>
<td>1.2</td>
<td>1.9</td>
<td>200 MHz</td>
<td>12 cycles</td>
<td>0.15 mm²</td>
<td>60 μW/MHz</td>
</tr>
</tbody>
</table>

*: 2.29 in ARM mode but 2.21 in Thumb2 mode (16-32bit Instructions)
More efficiency: Heterogeneous Cluster

P2012 Cluster Architecture

Local Interconnect (Stream flow)

Shared Tightly Coupled Data Memory (TCDM)

Global Interconnect Interface

Debug and Test Unit (DTU)
HW-SW Interaction in the heterogeneous cluster

Data streaming Load/Store

Cluster Control

Hardware Processing

Data Mover

L3 Memory

Streaming data flow
Configuration and control flow

PEDF MOC
Memory (Communication) wall

Heterogeneous, Memory-exposed 3D Architecture
P2012 Fabric Memory Map

- Up to 32 clusters ➔ 544 processors
- L1 Memory ➔ up to 4 Mbyte/cluster; today 256 KB/cluster
- L2 Memory ➔ up to 12 Mb; today 1 MB
- GAS-NUMA architecture
- No caches!
Interconnect Hierarchy

Cluster Interconnect $\rightarrow$ 2 cycles

Host

Off-chip interconnect $\rightarrow$ 200 cycles

Global + Accelerator Interconnect $\rightarrow$ 20 cycles

L3 (DRAM)
Cluster interconnect
Features:
- 2/3 cycles latency
- 0/1 pipe stalls
- Non-blocking
- Parametric soft IP
L1 ultra-low latency interconnect

Ultra-low latency $\rightarrow$ short wires + synchronicity

World-level bank interleaving «emulates» multiported mem
Shared Peripherals

Processors

Routing Tree

Arbitration Tree

Peripherals
Banking conflicts?

![Graph showing shared memory conflicts with theoretical and measured computation times.](image)

**Typical case**

- LD/ST %: 0%, 35.70%, 50%, 66.70%, 100%
- Computation Time: Theoretical vs Measured
- Values: 1, 1.0328, 1.0721, 1.1329, 1.2536
Implementation challenges

Clock borrowing (skewing) → tool supports it
Really helpful!

Global Wirelength minimization → manual effort!
GR-aware LogICo design needed

Timing-driven floorplanning → manual effort
Tool runtime is a major issue

Multi-corner optimization → supported but…
Huge P, T differences between corners.
Meeting $T_{max}$ WC makes a pretty bad AC netlist!

Memory cut optimization → manual effort!
Multi-dimensional: speed (setup, hold), leakage, density….
Inter-cluster interconnect
Cluster Decoupling

- Each cluster has its own operating point tunable to the right energy budget
- Inter cluster communication must support cluster decoupling
LETI’s ANoC Technology

NoC Architectures
- Communication: Network on Chip
  Application: Telecom MC-CDMA
  Static data flow programming model
- Dynamic reconfiguration
  SYSTEMC-TLM models
  ANoC architecture
- 3D NoC
  Flexible IPs
  Application: 3GPP-LTE & multimedia

Low-Power Adaptation
- VDD hopping
  Distributed control
- Fine grain DVFS
  GALS chips
  Memory usage
- Variability aware PWM
  Distributed optimization
  Hierarchical control

Technology
- First asynchronous library
- Optimized asynchronous lib.
  Asynchronous flow
  Low-power macros: FLL, Multiprobes, TFS, TFR
  3D flow, testability study, 3D serialization techniques

Timeline:
- 2003
- 2006
- 2008
- 2009
- 2011

Technology Versions:
- FAUST 130 nm
- ALPIN 65 nm
- MAGALI 65 nm
- Locomotiv 32 nm
- MAG3D 65 nm

P2012 Integration

With P2012
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
ANoC router

- **Wormhole source routing**
  - Path is in the header flit
  - Supports any topology
- **Possible virtual channels**
  - Allows QoS
    - Low-latency
    - Best-effort
- **Modular micro-architecture**
  - Several input controllers
    - Handle routing
  - Several output controllers
    - Handle arbitration

![Diagram of ANoC router with header and payload structures and possible virtual channels.]
P2012 SoC GANoC

Topology is tailored to architecture

System plug: asymmetric bandwidth

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

- SERDES (adapt to SoC NoC IF bitwidth)
- Protocol translation (possible multiple stages ANoC→SNOC→STBUS→AXI)
- Configurable address remapping (more later)
- Sideband signals adaptation
LD/ST and DMA memory transfers

- **Intra-Cluster:**
  - LD/ST (UMA)
  - DMA: From/to TCDM to/from HWPE

- **Inter-Cluster:**
  - LD/ST (NUMA)
  - DMA: L1-to/from-L1

- **Cluster to/from L2-Mem:**
  - LD/ST (NUMA)
  - DMA: L1 to/from L2

- **Cluster to/from L3-Mem (though the system bridge):**
  - LD/ST (NUMA)
  - DMA: L1 to/from L3

**DMA BW 6.4GBps/Cluster → external memory bandwidth?**
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
Reducing TSV cost

- 3D links is up to 4 times faster than 2D links
- Reducing area is thus possible without bandwidth loss
- MD = Medium density = 10 µm diameter
- HD = High density = 1 µm diameter

10 X
Advanced 3D Integration

- Main memory → TSV-based DRAM wide-IO
- On-chip memory → 3D-NoC to active interposer
June 2012 ....

- Fully functionnal
- High yield
- Actual perf higher than expected

Currently not in STHORM product roadmap
A Software view
Breaking the Memory wall: Memory-exposed Parallel Programming

OpenCL memory model

- Compute Unit 1
  - Private
  - Work Item
  - Local Memory
- Compute Unit 2
  - Private
  - Work Item
  - Local Memory

Global Memory
Programming Heterogeneous Parallelism

Task Parallelism

Data Parallelism

OpenCL

ARM Cortex-A15

P2012

Rogue

Multi-core CPU

Many-core

GPU

More parallelism

More programmability
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!
Key issue: Sharing large data buffers
How to efficiently share?

- **Split L3 (PC legacy) → inefficient! Lots of copying**

- **Unified L3… devil is in details**
  - Caching
  - Virtualization
  - QoS

AMD fusion APU Zacate (90GFLOP, 15WDTP)
Accelerator coherency (AXI ACP)

- Asymmetric: P2012 accesses ARM L1
- Physically addressed & tagged (no virtualization)
- Not always beneficial → cache pollution, traffic
Current AXI4 solution is NOT scalable
Conclusions

- Not one, but many interconnects
  - Cluster → latency + speed challenge
  - Fabric → latency, bandwidth & decoupling
  - System → SW abstractions, protocol conversion, L3 bandwidth
- Lots of nice ideas to explore
  - Cluster: speed adaptation via pipe stage insertion
  - Fabric: shorten latency, fast synchronization support
  - System: scalable virtualization & SW-driven coherency, multicast, interleaving…
- NoCs are increasingly critical!