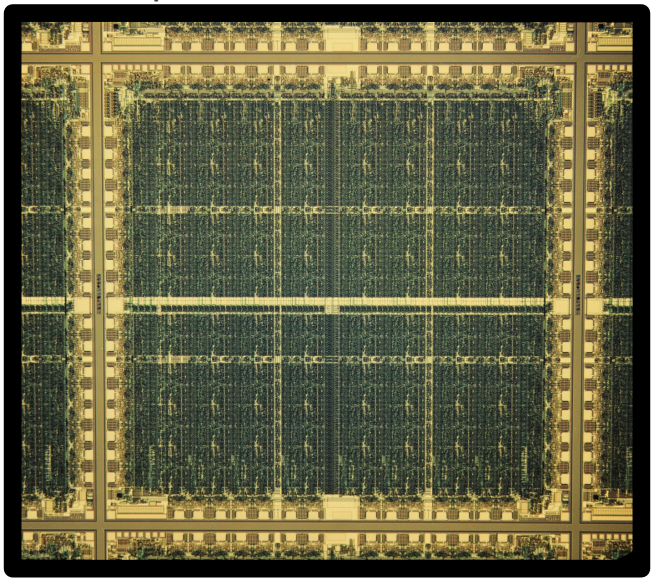
XILINX
ALL PROGRAMMABLE™

From modern FPGA to high-level post-modern C++ abstractions for heterogeneous computing with OpenCL SYCL & SPIR-V

Ronan Keryell, Xilinx Research Labs & Khronos OpenCL & SYCL C++ committee

HiPEAC WRC 2016/01/19

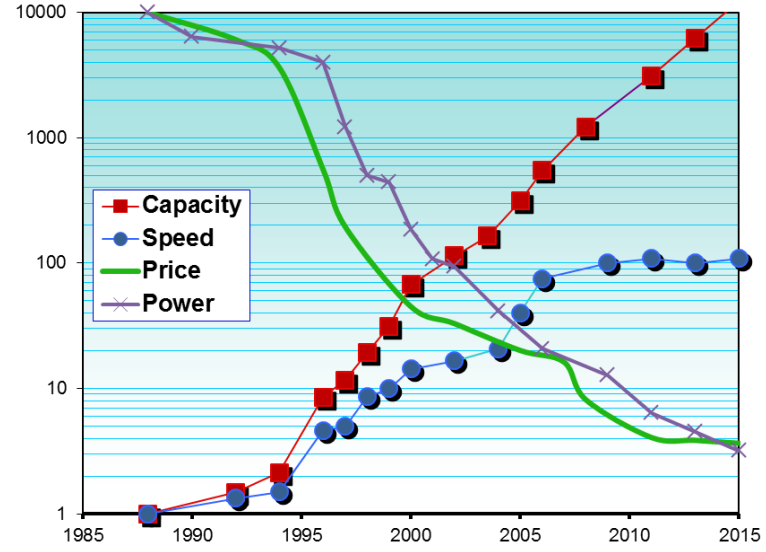
Once upon a time the XC2064…



1985: the First FPGA

- **64 flip-flops**
- **128 3-LUT**
- **58 I/O pins**
- **18 MHz (toggle)**
- **2 µm 2LM**

Since then…



- **10,000x more logic…**
  - Plus embedded IP
    - Memory
    - Microprocessor
    - DSP
    - Gigabit Serial I/O
- **100x faster**
- **5,000x lower power/gate**
- **10,000x lower cost/gate**

Three Ages of the FPGAs: Trimberger, S, Proceedings of the IEEE | Vol. 103, No. 3, March 2015

Next generation challenges

- Power
- Performance
- Cost drivers
- Power management
- 64bit processing
- Real-time processing
- Video and graphics processing
- Pervasive safety and security
- Higher levels of processor-fabric integration

Introducing the Zynq UltraScale+ MPSoC

ARM Cortex A53 & R5

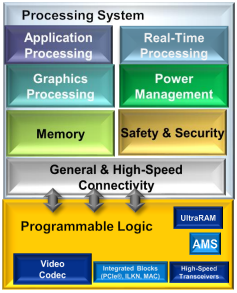
- Application processing subsystem
- Real Time processing subsystem
- Memory subsystem for max performance

IO, Video, Graphics

- Next-generation coherent interconnect
- High-speed I/Os (PCIE,USB3,SATA,GbE)
- Graphics and Video Processing Engines

Advanced Power Mgmt

- Fine-grained power reduction
- System-level software & run time opt



Security & Safety

- Enhanced Authentication, Encryption, Antitamper and trust
- Safety with industry standards support

Fabric Acceleration

- UltraScale+ fabric with time borrow
- FinFET performance and power
- HD UlltraRAM, and enhanced DSP

XCVRs & Protocols

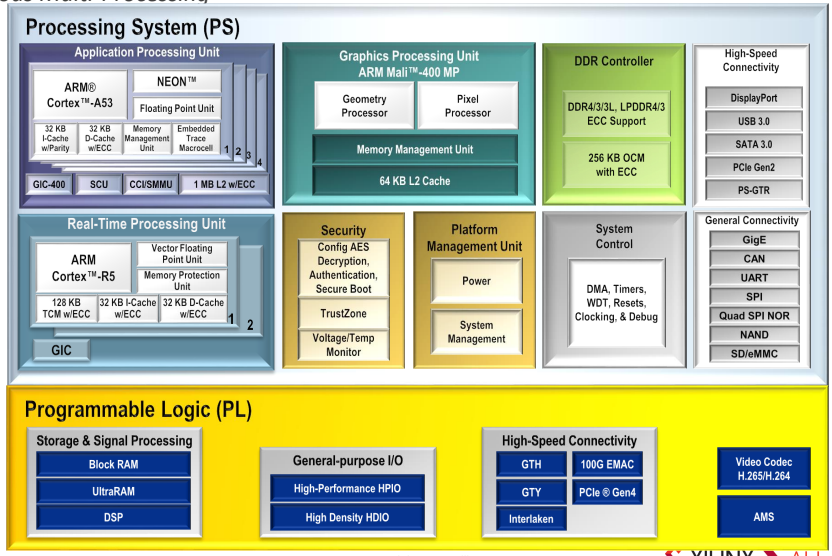
- Power efficient, 32Gbps
- 100G Ethernet and 150G Interlaken
- PCIe Gen3 & Gen4

Software & Tools

Run Time (Xilinx)	Run Time (Ecosystem)	Tools
- Linux (64b)	- FreeRTOS	- Xilinx SDK
- Hypervisor	- Micrium	- Vivado®
- OpenAMP	- WindRiver & More	- SDx environments

Zynq UltraScale+ MPSoC Overview

Heterogeneous Multi-Processing



Introducing the Zynq UltraScale+ MPSoC

ARM Cortex A53 & R5

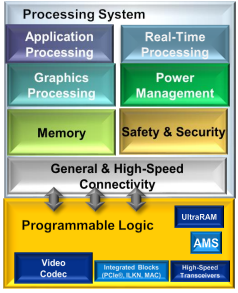
- Application processing subsystem
- Real Time processing subsystem
- Memory subsystem for max performance

IO, Video, Graphics

- Next-generation coherent interconnect
- High-speed I/Os (PCIE,USB3,SATA,GbE)
- Graphics and Video Processing Engines

Advanced Power Mgmt

- Fine-grained power reduction
- System-level software & run time opt



Security & Safety

- Enhanced Authentication, Encryption, Antitamper and trust
- Safety with industry standards support

Fabric Acceleration

- UltraScale+ fabric with time borrow
- FinFET performance and power
- HD UlltraRAM, and enhanced DSP

XCVRs & Protocols

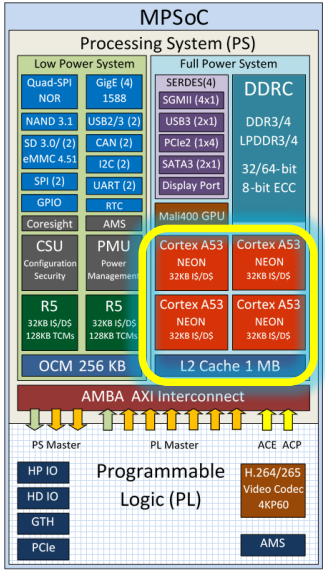
- Power efficient, 32Gbps
- 100G Ethernet and 150G Interlaken
- PCIe Gen3 & Gen4

Software & Tools

Run Time (Xilinx)	Run Time (Ecosystem)	Tools
- Linux (64b)	- FreeRTOS	- Xilinx SDK
- Hypervisor	- Micrium	- Vivado®
- OpenAMP	- WindRiver & More	- SDx environments

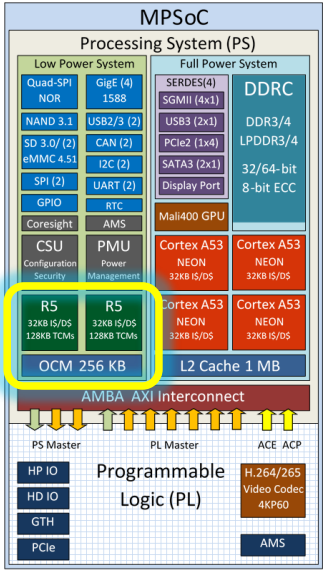
Application Processing Subsystem

- **Quad Cortex-A53 64-bit CPU**
  - 32KB each of L1 I & D$ with ECC/Parity
  - 1 MB L2 Cache with ECC
  - Virtualization Support
  - Crypto instructions support
- **IO Coherency with ACP & ACE-Lite**
- **Full Coherency between APU & PL**
- **Up to 1.5 GHz Frequency**
- **Power-gating**
  - Per core power-gating
  - L2 power-gating



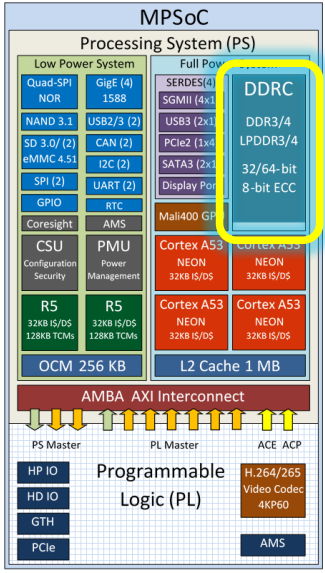
Real-time Processing Subsystem

- **Dual Core Cortex-R5 RPU**
  - Cortex-R5 Lockstep
  - Single & Double precision FPU
  - 32KB of L1 I & D caches w/ ECC
  - 256KB TCM with ECC
- **OCM (On-Chip-Memory)**
  - 256KB OCM with ECC
  - AXI Exclusive monitor support
  - Can be partitioned between different subsystems
- **Up to 600MHz Frequency**
- **Low power domain**
  - Full power domain completely powered off
  - R5s & USBs power-gate-able

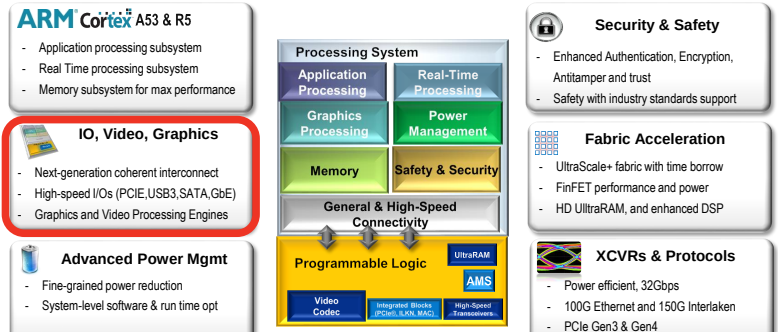


Memory Subsystem

- **Six-port DDR Controller**
  - Supports exclusive monitors
- **32 or 64-bit DDR with ECC**
- **DDR3/4 and LPDDR3/4**
- **Up to 2400 Mb/s/pin**
- **QoS support for 3 traffic classes**
  - Low latency, Real-time, Best-effort
  - Guaranteed latency for RT
- **Memory protection, partitioning, and TrustZone support**
  - Using XMPU

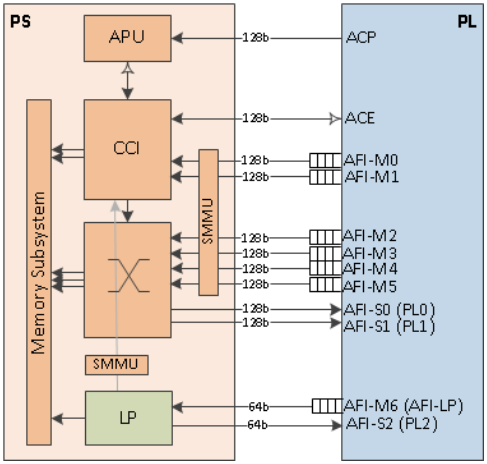


Introducing the Zynq UltraScale+ MPSoC



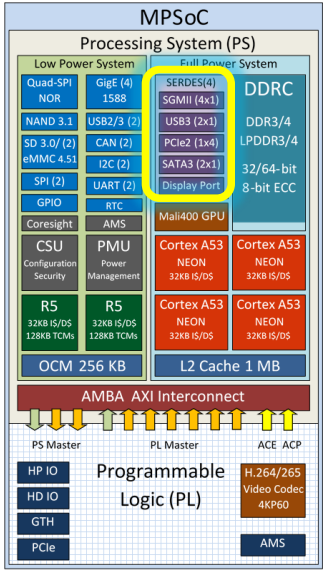
PS ⇔ PL “Data Mover” Interfaces

- **PL master ports**
  - AFI master ports
  - PL IO-coherency via CCI
  - PL virtualization via SMMU
- **PL Slave Ports**
  - PS-to-PL data movers
  - Memory mapped
- **ACP (Accelerated Coherency Port)**
  - ACP for IO (one-way) coherency
- **ACE (AXI Coherency Extensions)**
  - Full coherency between PS & PL
- **Per-port bandwidth = 85 Gb/s**
  - Read+write bandwidth



High Speed I/Os

- **USB2/3**
  - 2 independent controllers
  - OTG, Host, Device
- **SATA3**
  - Up to 2 channels
- **Display Port**
  - 4KP30 support
  - 1-2 lanes
- **PCIe Gen2 Rootport or Endpoint**
  - PCIe Gen3/4 EP also hardened
- **SGMII for GbE**
  - 4 independent GbE controllers
- **Tightly integrated transceivers**



Dedicated Video Processing Engines

Graphics Processing Unit, Video CODEC & DisplayPort

Video CODEC Unit (VCU)

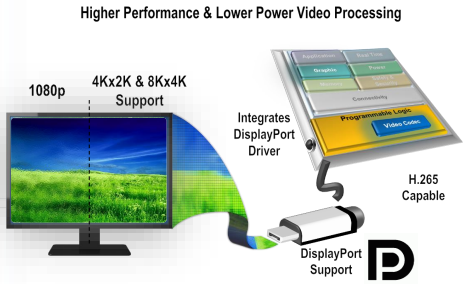
- More efficent vs. software implementation
  - Higher display density, faster encoding
  - Lower power consumption
- H.265 (HEVC) 8Kx4K (15 fps) 4Kx2K (60 fps)
- 8 and 10 bit per color component
- I, P, B frame support for highest compression

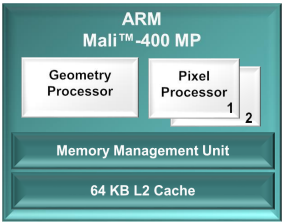
DisplayPort

- Video resolution up to 4Kx2K (30 fps)
- Audio up to 8 channels of 24-bit at up to 192 KHz
- Reducing BOM cost by eliminating display driver

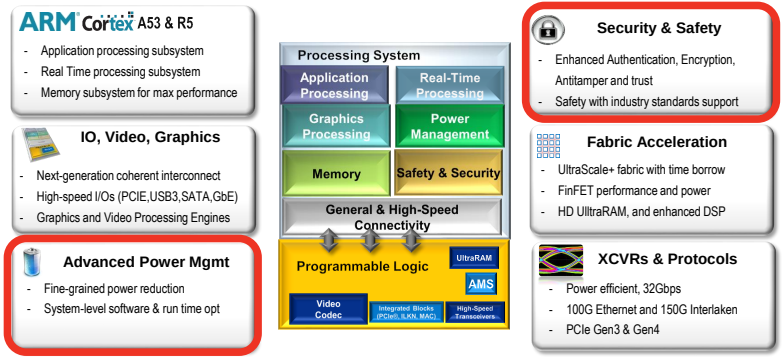
Graphics Processing Unit (GPU)

- 3D visual, HMI, instrumentation, waveform display
- 1080p resolution graphics
- Mali-400 MP2 up to 667 MHz frequency (OpenGL ES 2.0, 13 GFLOPS)



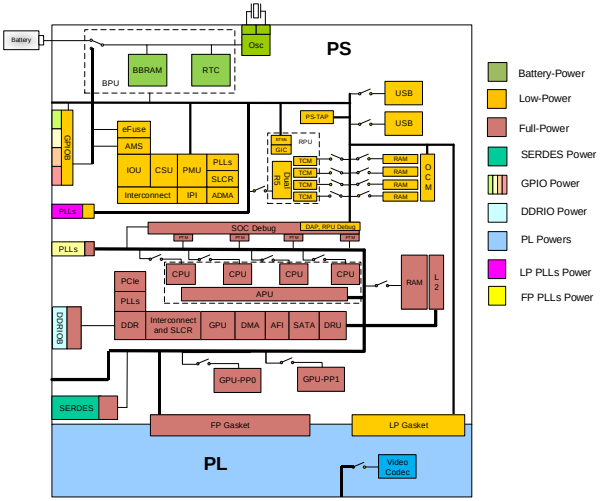


Introducing the Zynq UltraScale+ MPSoC



Power-Domains and Power-Gating

- **Multiple power domains**
  - Low power domain
  - Full power domain
  - PL power domain
- **Power gating**
  - A53 per core
  - L2 and OCM RAM
  - GPU, USB
  - R5s & TCM
  - Video CODEC
- **Sleep Mode**
  - 35mW sleep mode
  - Suspend to DDR with power off



Security, Safety & Reliability

Advanced Device-Level Secure Processing

- Information Assurance, Anti-Tamper, Trust
- Multi-layered Authentication for Secure System Boot
- Key Management & Revocation

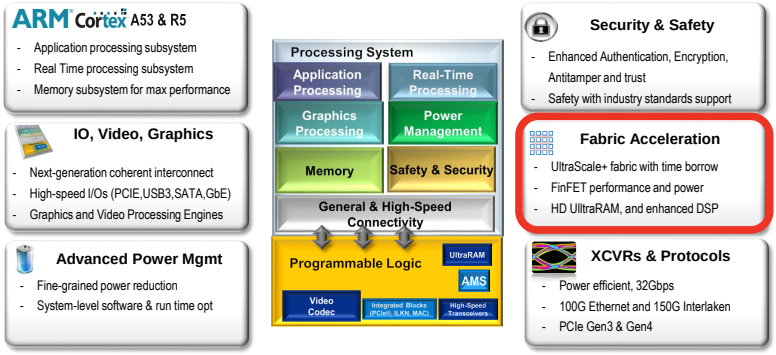
Architected for Safe Systems

- IEC61508 & ISO26262 Functional Safety Standards
- Redundancy, Diversity and Lock-step
- Layered Partitioning: Core / Infrastructure / Peripherals

Delivering High Reliability

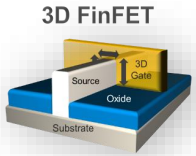
- High Availability Systems
- Error Detection & Handling
- Subsystem Isolation & Protection

Introducing the Zynq UltraScale+ MPSoC



Tuned Process for Optimal Performance/Watt

Optimal Operating Voltage Selection



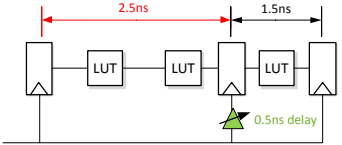
3D Gate "wraps" around channel for more surface area, achieving

- ✓ Faster transistor on/off switching speeds for greater performance
- ✓ Lower leakage and operating voltage for lower power

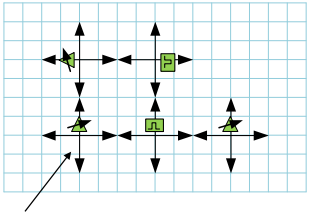
Tuned Operating Voltage (V_{CCINT})	1V	0.95V	0.85V	0.72V
	7 Series (28nm) V_{NOM}	20% Greater Performance → UltraScale (20nm) V_{NOM}	33% Greater Performance → UltraScale+ (16nm) V_{NOM}	UltraScale+ (16nm) V_{LOW}
Normalized Fabric Performance	1.0x	1.2x	1.6x	1.2x
Normalized Total Power	1.0x	.7x	.8x	.5x
Performance/Watt	1.0x	1.7x	2x	2.4x

Time Borrow in the Fabric

- **Time-borrowing concept**
  - Shift available slack from fast stages to performance-critical paths

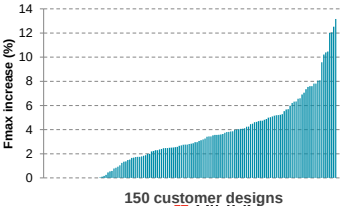


Example	Tmin	Fmax
Baseline	2.5 ns	400 MHz
Time Borrow	2 ns	500 MHz



- **UltraScale+ time-borrowing platform**
  - Fine-grain delays to adjust clock skew
  - Programmable pulse generators for latch-based time borrow

- **High Performance without design changes**
  - Very effective on high-performance designs
  - Transparent to customers, part of default flow



UltraRAM: New Memory Technology

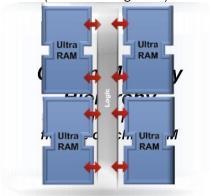
Distributed RAM
(bits to kilobits)

- Wide, shallow FIFOs
- Shift registers
- State machines

Block RAM
(10s of megabits)

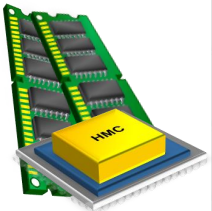
- Data/coefficient storage
- Deep FIFOs
- Shallow buffering

UltraRAM
(100s of megabits)



- Deep packet buffering
- Video buffering
- State, statistics, counters

External Memory
(100s of megabits to gigabits)



Larger data storage

Up to 432 Mb to replace external memory for cost, power, performance

Introducing the Zynq UltraScale+ MPSoC

ARM Cortex A53 & R5

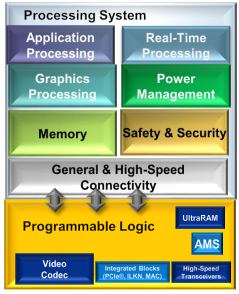
- Application processing subsystem
- Real Time processing subsystem
- Memory subsystem for max performance

IO, Video, Graphics

- Next-generation coherent interconnect
- High-speed I/Os (PCIE,USB3,SATA,GbE)
- Graphics and Video Processing Engines

Advanced Power Mgmt

- Fine-grained power reduction
- System-level software & run time opt



Security & Safety

- Enhanced Authentication, Encryption, Antitamper and trust
- Safety with industry standards support

Fabric Acceleration

- UltraScale+ fabric with time borrow
- FinFET performance and power
- HD UlltraRAM, and enhanced DSP

XCVRs & Protocols

- Power efficient, 32Gbps
- 100G Ethernet and 150G Interlaken
- PCIe Gen3 & Gen4

Software & Tools

Run Time (Xilinx)	Run Time (Ecosystem)	Tools
- Linux (64b)	- FreeRTOS	- Xilinx SDK
- Hypervisor	- Micrium	- Vivado®
- OpenAMP	- WindRiver & More	- SDx environments

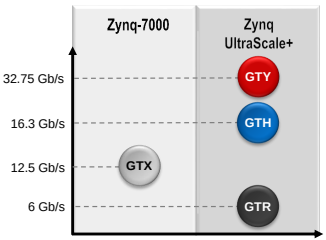
Enhanced transceivers

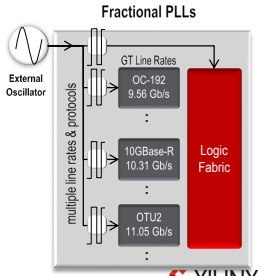
Diverse, Power Efficient SerDes for Bandwidth

- 16G (GTH) & 32G (GTY) transceivers in PL,
- 6G (GTR) in PS for direct access to key processing elements, with full PHY/IP compliance for key protocols:
  - USB, SATA, DisplayPort, PCIe, Ethernet

Fractional PLLs to Reduce BOM Cost

- Single external oscillator generates GT & logic fabric clocks for multiple non-integer line rates
- Available in GTH, and GTY transceivers





Introducing the Zynq UltraScale+ MPSoC

ARM Cortex A53 & R5

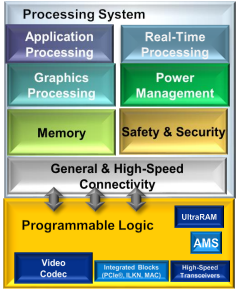
- Application processing subsystem
- Real Time processing subsystem
- Memory subsystem for max performance

IO, Video, Graphics

- Next-generation coherent interconnect
- High-speed I/Os (PCIE,USB3,SATA,GbE)
- Graphics and Video Processing Engines

Advanced Power Mgmt

- Fine-grained power reduction
- System-level software & run time opt



Security & Safety

- Enhanced Authentication, Encryption, Antitamper and trust
- Safety with industry standards support

Fabric Acceleration

- UltraScale+ fabric with time borrow
- FinFET performance and power
- HD UlltraRAM, and enhanced DSP

XCVRs & Protocols

- Power efficient, 32Gbps
- 100G Ethernet and 150G Interlaken
- PCIe Gen3 & Gen4

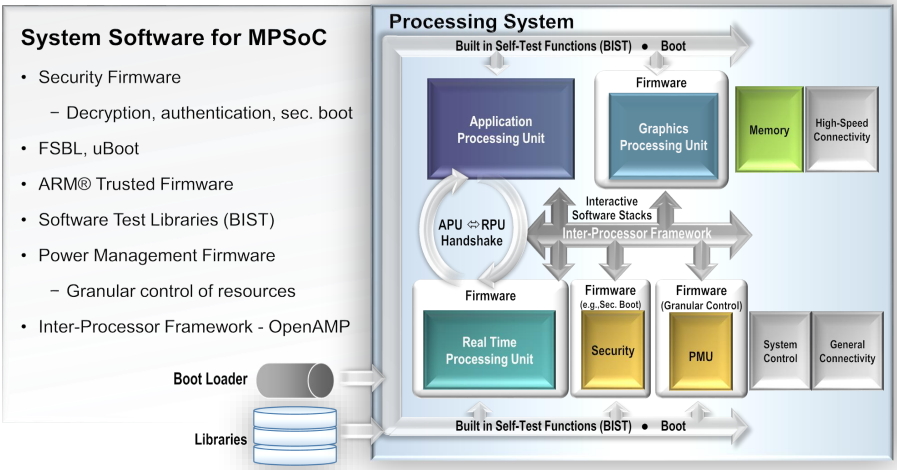
Software & Tools

Run Time (Xilinx)	Run Time (Ecosystem)	Tools
- Linux (64b)	- FreeRTOS	- Xilinx SDK
- Hypervisor	- Micrium	- Vivado®
- OpenAMP	- WindRiver & More	- SDx environments

System Software

System Software for MPSoC

- Security Firmware
  - Decryption, authentication, sec. boot
- FSBL, uBoot
- ARM® Trusted Firmware
- Software Test Libraries (BIST)
- Power Management Firmware
  - Granular control of resources
- Inter-Processor Framework - OpenAMP

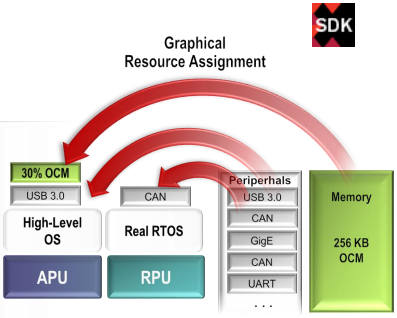


Xilinx SDK

Integrated Environment for *Software Design*

New & Enhanced Software Tools

- Heterogeneous Multicore Debug
  - Debug & cross triggering for APU/RPU/MicroBlaze™ Processor
- System-level Profiling and Performance Analysis Tools
  - Analysis for interfaces
  - Across processing & Programmable Logic (PL) domains
- Multi-OS Boot Image Tool
  - Creates boot image(s)
  - Supports output from Xilinx & 3rd party IDEs
- System Resource Partitioning Tool
  - Graphic assignment of resources



Industry-standard Tools Support

- Enabling developer-preferred dev /debug environments

YOKOGAWA

ARM DS Development Tools

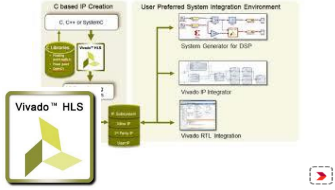
LAUTERBACH DEVELOPMENT TOOLS

KMC

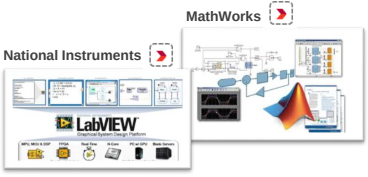
Enabling Smarter Systems

Abstractions

Software Automation



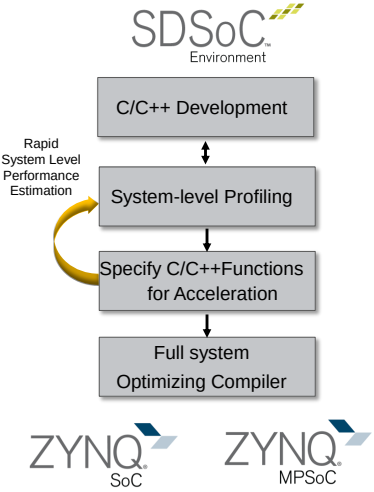
Hardware Automation



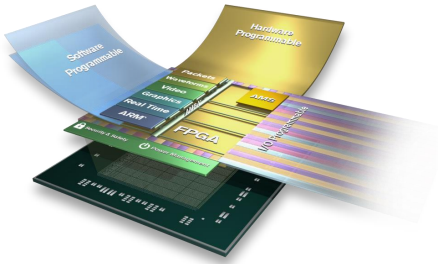
System Automation

SDSoC Development Environment

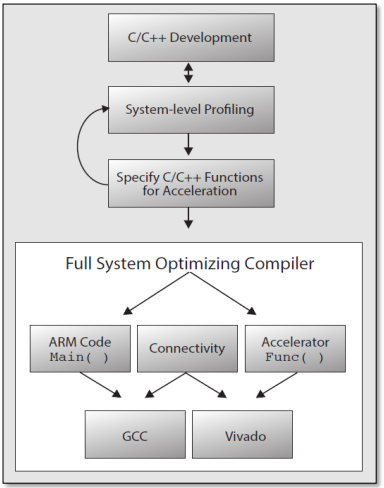
C/C++ Programming for Zynq SoC and MPSoC



- ASSP-like programming experience
- System-level profiling
- Full system optimizing compiler
- Expert use model for platform developers and system architects

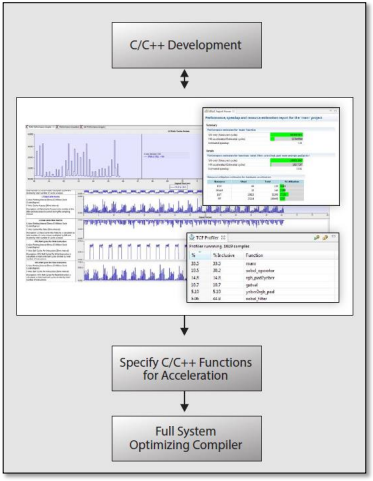


SDSoC: Full System Optimizing Compiler



- **Automated Connectivity Optimization**
  - Finds the Data Mover and PS-PL interface for optimal dataflow
  - Rapid exploration of different system connectivity topologies
- **Rapid Software Configurable Application Acceleration using C/C++**
  - Automated function acceleration in programmable logic
  - Up to 100X increase in performance vs. software
  - System optimized for latency, bandwidth, and hardware utilization

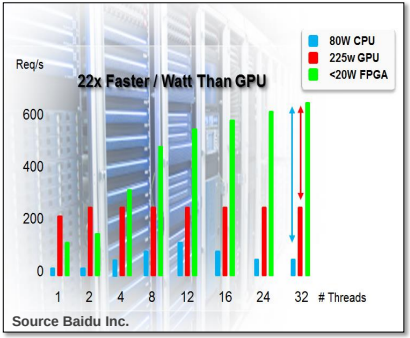
SDSoC: System Level Profiling



- **Rapid system performance estimation**
  - Full system estimation (programmable logic, data communication, processing system)
  - Reports SW/HW cycle level performance and hardware utilization
- **Automated performance measurement**
  - Runtime measurement by instrumentation of cache, memory, and bus utilization

SDAccel: Development Environment

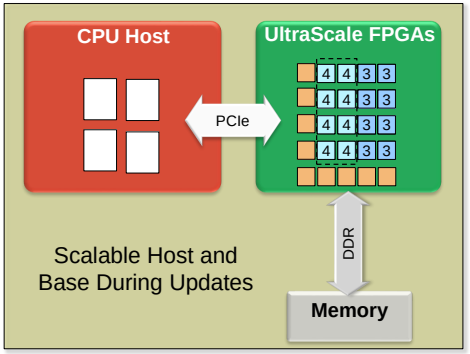
Replace CPU/GPU with up to 25X performance/watt value of FPGAs



Ideal for data center, A&D and medical imaging accelerated computing

- **CPU/GPU like development experience**
  - OpenCL, C/C++ Kernel acceleration
  - Khronos OpenCL conformant
- **Architecturally optimizing compiler**
  - 25x Performance/Watt advantage over GPUs / CPUs
- **Complete CPU/GPU like run-time experience**
  - Run-time reconfigurable accelerators
  - Production ready platforms

CPU/GPU-like Runtime Experience on FPGAs

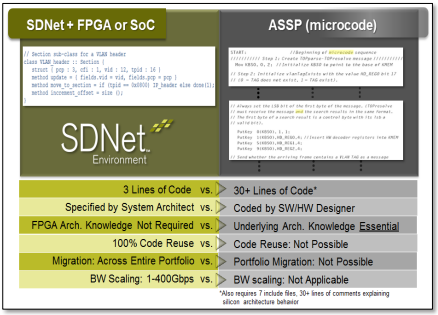


CPU/GPU Runtime Experience

- On-demand loadable acceleration units
- Always on interfaces (Memory, Ethernet PCIe, Video)
- Optimize resources thru hardware reuse

SDNet: SW Defined Specification Environment

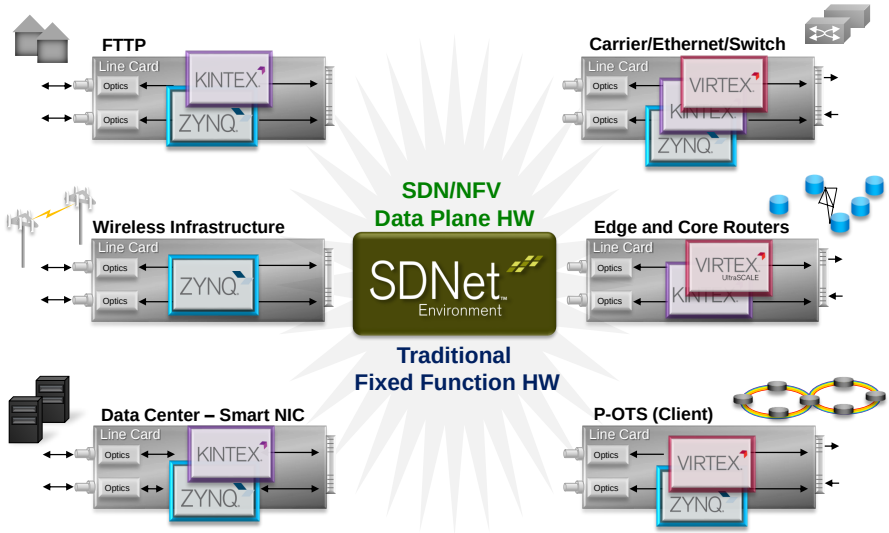
Replaces NPUs with All Programmable performance, flexibility, and security



10x productivity advantage

- **‘Softly Defined’ networks**
  - Content intelligence data plane hardware supporting hit-less in service updates
  - Dynamically collaborates with control plane
  - Address performance, flexibility, and security challenges of agile service-oriented networking
- **SDNet’s specification driven environment**
  - Generates packet data plane HW subsystem
  - Generates subsystem firmware
  - Generates optimized packet processing engines and flows
    - e.g. parsing, editing, search, and feature optimized QoS policy
  - Generates test benches

Data Plane Function Acceleration – Core to Edge



Can we get higher level-programming in some standard language?

- FPGA $\equiv$ MPSoC with amazing features
- $\exists$ Huge libraries waiting for acceleration on them...
- Computation-intensive libraries often written in C/C++/Fortran (TernsorFlow (Google...), Caffe (Berkeley, Yahoo...), Torch7 (Facebook...) for DNN...)
  - Often with OpenMP or CUDA single-source extensions
  - Often with vector intrinsics or assembly code
- Need fine control of real hardware resources for performance & power efficiency

Outline

C++14

- 2 Open Source compilers available *before* ratification (GCC & Clang/LLVM)
- Strategy: 1 major (C++11) and 1 minor (C++14) version on a 6-year cycle
- Next big version expected in 2017 (C++1z)
  - Already being implemented! ☺
- Well defined memory model for parallel programming!
- Monolithic committee replaced by many smaller *parallel* task forces
  - Parallelism TS (Technical Specification) with Parallel STL
  - Concurrency TS (threads, mutex...)
  - Array TS (multidimensional arrays *à la* Fortran)
  - Transactional Memory TS...

Race to parallelism! Definitely matters for HPC and heterogeneous computing!

C++ is a complete new language

- Forget about C++98, C++03...
- Send your proposals and get involved in C++ committee (pushing heterogeneous computing)!

Modern C++ & HPC *(I)*

- Huge library improvements
  - <thread> library and multithread memory model <atomic> ⇝ HPC
  - Hash-map
  - Algorithms
  - Random numbers
  - ...
- Uniform initialization and range-based for loop

```
std::vector<int> my_vector { 1, 2, 3, 4, 5 };
for (int &e : my_vector)
  e += 1;
```

- Easy functional programming style with λ expressions (anonymous functions)

```
std::transform(std::begin(v), std::end(v), [] (int e) { return 2*e; });
```

Modern C++ & HPC *(II)*

- Lot of meta-programming improvements to make meta-programming ~~easy~~ easier: variadic templates, type traits <type_traits>...
- Make simple things simpler to be able to write generic numerical libraries, etc.
- Automatic type inference for terse programming
  - Python 3.x (interpreted):

```
def add(x, y):
  return x + y
print(add(2, 3))      # 5
print(add("2", "3")) # 23
```

  - Same in C++14 but compiled + static compile-time type-checking:

```
auto add = [] (auto x, auto y) { return x + y; };
std::cout << add(2, 3) << std::endl;        // 5
std::cout << add("2"s, "3"s) << std::endl; // 23
```

Without using templated code! ~~`template <typename >`~~ ☺

Modern C++ & HPC *(III)*

- R-value references & `std::move` semantics
  - matrix_A = matrix_B + matrix_C
    - Avoid copying (TB, PB, EB... ☹) when assigning or function return
- Avoid raw pointers, `malloc()`/`free()`/`delete[]`: use references and smart pointers instead

```
// Allocate a double with new() and wrap it in a smart pointer
auto gen() { return std::make_shared<double> { 3.14 }; }
[...]
{
  auto p = gen(), q = p;
  *q = 2.718;
  // Out of scope, no longer use of the memory: deallocation happens here
}
```

Modern C++ & HPC *(IV)*

- C++14 generalizes constexpr to statements

```
constexpr auto fibonacci(int v) {
  long long int u_n_minus_1 = 0;
  auto u_n = u_n_minus_1 + 1;
  for (int i = 1; i < v; ++i) {
    auto tmp = u_n;
    u_n += u_n_minus_1;
    u_n_minus_1 = tmp;
  }
  return u_n;
}
int main() {
  constexpr auto result = fibonacci(80);
  std::cout << result << std::endl;
  return 0;
}
```

Modern C++ & HPC *(V)*

Compiled to

```
movabsq $23416728348467685, %rsi # imm = 0x533163EF0321E5
movl    $_ZSt4cout, %edi
callq   _ZNSo9_M_insertIxEERSoT_
```

- Lot of other amazing stuff...
- Allow both low-level & high-level programming...
  - Great for heterogeneous computing!

C++11 std::thread

- It's all in the standard! http://en.cppreference.com/w/cpp/thread/thread

```
#include <chrono>
#include <iostream>
#include <thread>

void f(int n) {
  std::cout << "Thread " << n << "executing" << std::endl;
  std::this_thread::sleep_for(std::chrono::milliseconds(100));
}

  // Launch f in thread t:
  std::thread t(f, 1);
  // The same with a lambda
  std::thread t2([] {std::cout << "Hello!" << std::endl;});
```

- ∃ higher-level constructs: std::async, std::future/std::promise

Position argument

Which language for unified heterogeneous computing?

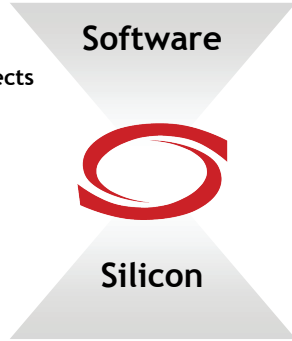
-  Entry cost
- ∃ thousands of dead parallel languages...
  -  Exit cost
- Use standard solutions with open source implementations
- Start with modern C++
  - Very successful & ubiquitous language
  - Very active since C++11 (C++14, C++1z...)
  - Interoperability: seamless interaction with embedded world, libraries, OS...
  - Combine both low-level aspects with high-level programming
    - Pay only for what you need (garbage collector or not...)
  - Classes can be used to define Domain Specific Embedded Language (DSEL)
  - Not directly targeting FPGA...
    - But extensible through classes (↝ DSEL)
    - Extensible with #pragma (already in Vivado HLS, SDSoC & SDAccel) and attributes

Outline

Khronos Connects Software to Silicon

Open Consortium creating OPEN STANDARD APIs for hardware acceleration
Any company is welcome – many international members – one company one vote

ROYALTY-FREE specifications
State-of-the art IP framework protects members AND the standards



International, non-profit organization
Membership fees cover operating and engineering expenses

Low-level silicon APIs needed on every platform
Graphics, parallel compute, rich media, vision, sensor and camera processing

API Specifications AND Conformance Tests for cross-vendor portability

Strong industry momentum
100s of man years invested by industry experts

Well over a *BILLION* people use Khronos APIs *Every Day…*

Interoperability nightmare in heterogeneous computing & graphics

- ∃ Many programming languages for heterogeneous computing
  - Writing compiler front-end may not be *the* real value for a hardware vendor...
    - Writing a C++1z compiler from scratch is almost impossible...
- ∃ Many programming languages for writing shaders
- Convergence in computing (Compute Unit) & graphics (Shader) architectures
  - Same front-end & middle-end compiler optimizations
- Need for some non source-readable portable code for IP protection

⇝ Defining common low-level representation !

SPIR-V transforms the language ecosystem

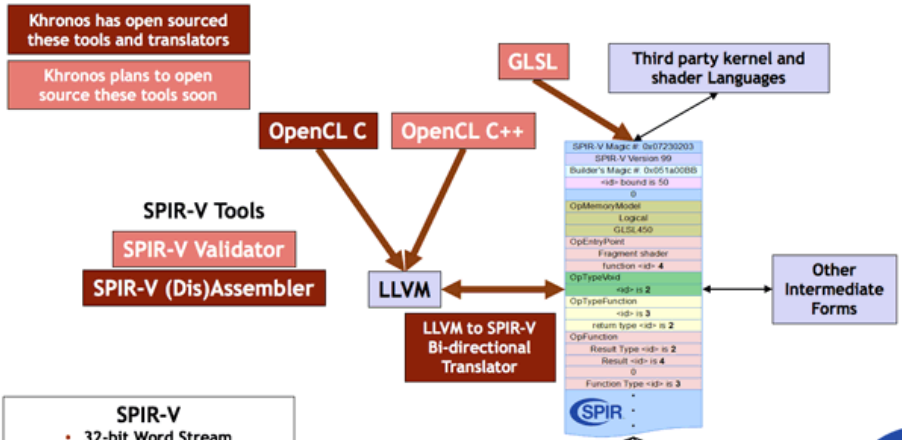
- First multi-API, intermediate language for parallel compute *and* graphics
  - Native representation for Vulkan shader and OpenCL kernel source languages
  - `https://www.khronos.org/registry/spir-v/papers/WhitePaper.pdf`
- Cross-vendor intermediate representation
  - Language front-ends can easily access multiple hardware run-times
  - Acceleration hardware can leverage multiple language front-ends
  - Encourages tools for program analysis and optimization in SPIR form

Evolution of SPIR family

SPIR	SPIR 1.2	SPIR 2.0	SPIR-V 1.0
LLVM Interaction	Uses LLVM 3.2	Uses LLVM 3.4	100% Khronos defined Round-trip lossless conversion
Compute Constructs	Metadata/Intrinsics	Metadata/Intrinsics	Native
Graphics Constructs	No	No	Native
Supported Language Feature Sets	OpenCL C 1.2	OpenCL C 1.2 OpenCL C 2.0	OpenCL C 1.2 / 2.0 OpenCL C++ and GLSL
OpenCL Ingestion	OpenCL 1.2 Extension	OpenCL 2.0 Extension	OpenCL 2.1 Core OpenCL 1.2 / 2.0 Extensions
Vulkan Ingestion	-	-	Vulkan 1.0 Core

Not based on LLVM to isolate from LLVM roadmap changes

Driving SPIR-V Open Source ecosystem



SPIR-V 1.0 Resources

- SPIR-V 1.0 specification available in Khronos Registry
  `https://www.khronos.org/registry/spir-v`
- Feedback forum for questions and feedback
  `https://forums.khronos.org/showthread.php/12919-Feedback-SPIR-V`
- Whitepaper “An Introduction to SPIR-V”
  `https://www.khronos.org/registry/spir-v/papers/WhitePaper.html`
- Bug reporting
  `https://www.khronos.org/bugzilla/enter_bug.cgi?product=SPIR-V`
- SPIR-V tools project including an assembler, binary module parser, disassembler & validator for SPIR-V `https://github.com/KhronosGroup/SPIRV-Tools`
- LLVM framework with SPIR-V support including an LLVM↔SPIR-V bi-directional converter `https://github.com/KhronosGroup/SPIRV-LLVM`
- GLSL compiler in development

Missing link...

- No tool providing
  - Modern C++ environment
  - Heterogeneous computing
  - Single source for programming productivity
  - OpenCL interoperability

Outline

What about heterogenous computing???

- C++ std::thread is great...
- ...but supposed shared unified memory (SMP) ☹
  - What if accelerator with own separate memory?
  - What if using distributed memory multi-processor system (MPI...)?
- ⇝ Extend the concepts...
  - Replace raw unified-memory with buffer objects
  - Define with accessor objects which/how buffers are used
  - Since accessors are already here to define dependencies, no longer need for std::future/std::promise! ☺
  - Add concept of queue to express where to run the task
  - Also add all goodies for massively parallel accelerators (OpenCL/Vulkan/SPIR-V) in clean C++

SYCL ≡ pure C++14 DSEL

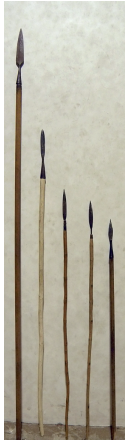
- Implement concepts useful for heterogeneous computing
- Asynchronous task graph
- Buffers to define location-independent storage
- Accessors to express usage for buffers and pipes: read/write/...
- Hierarchical parallelism
- Hierarchical storage
- Single source programming model
  - Take advantage of CUDA & OpenMP simplicity and power
  - Compiled for host *and* device(s)
  - Enabling the creation of C++ higher level programming models & C++ templated libraries
- Most modern C++ features available for OpenCL
  - Programming interface based on abstraction of OpenCL components (data management, error handling...)
  - Provide OpenCL interoperability
- Host fallback (debug and symmetry for SIMD/multithread on host)
- Directly executable DSEL
- Host emulation for free & no compiler needed for experimenting

Puns and pronunciation explained

OpenCL SYCL

sickle ['si-kəl]

OpenCL SPIR



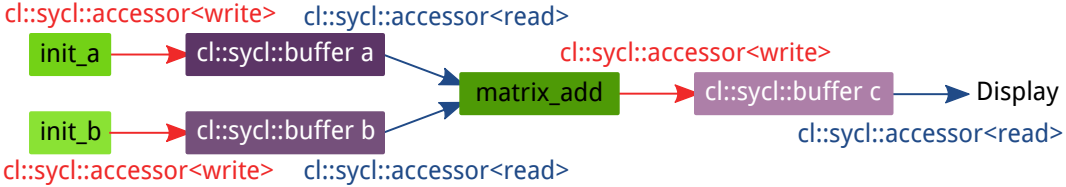
spear ['spir]

Complete example of matrix addition in OpenCL SYCL

```
#include <CL/sycl.hpp>
#include <iostream>

using namespace cl::sycl;

constexpr size_t N = 2;
constexpr size_t M = 3;
using Matrix = float[N][M];

int main() {
 Matrix a = { { 1, 2, 3 }, { 4, 5, 6 } };
 Matrix b = { { 2, 3, 4 }, { 5, 6, 7 } };

 Matrix c;

 {// Create a queue to work on default device
  queue myQueue;
  // Wrap some buffers around our data
  buffer<float, 2> A { a, range<2> { N, M } };
  buffer<float, 2> B { b, range<2> { N, M } };
  buffer<float, 2> C { c, range<2> { N, M } };
  // Enqueue some computation kernel task
  myQueue.submit([&](handler& cgh) {
   // Define the data used/produced
   auto ka = A.get_access<access::read>(cgh);
   auto kb = B.get_access<access::read>(cgh);
   auto kc = C.get_access<access::write>(cgh);
   // Create & call OpenCL kernel named "mat_add"
   cgh.parallel_for<class mat_add>(range<2> { N, M },
      [=](id<2> i) { kc[i] = ka[i] + kb[i]; }
   );
  }); // End of our commands for this queue
 } // End scope, so wait for the queue to complete.
   // Copy back the buffer data with RAII behaviour.
 return 0;
}
```

Asynchronous task graph model

- Change example with initialization kernels instead of host
- Theoretical graph of an application described *implicitly* with kernel tasks using buffers through accessors



- Possible schedule by SYCL runtime:

⇝ Automatic overlap of kernels & communications
- Even better when looping around in an application
- Assume it will be translated into pure OpenCL event graph
- Runtime uses as many threads & OpenCL queues as necessary (GPU synchronous queues, AMD compute rings, AMD DMA rings...)

Task graph programming — the code

```
#include <CL/sycl.hpp>
#include <iostream>
using namespace cl::sycl;
// Size of the matrices
constexpr size_t N = 2000;
constexpr size_t M = 3000;
int main() {
  { // By sticking all the SYCL work in a {} block, we ensure
    // all SYCL tasks must complete before exiting the block

    // Create a queue to work on
    queue myQueue;
    // Create some 2D buffers of float for our matrices
    buffer<double, 2> a({ N, M });
    buffer<double, 2> b({ N, M });
    buffer<double, 2> c({ N, M });
    // Launch a first asynchronous kernel to initialize a
    myQueue.submit([&](auto &cgh) {
        // The kernel write a, so get a write accessor on it
        auto A = a.get_access<access::write>(cgh);

        // Enqueue parallel kernel on a N*M 2D iteration space
        cgh.parallel_for<class init_a>({ N, M },
                                       [=] (auto index) {
                                         A[index] = index[0]*2 + index[1];
                                       });
      });
    // Launch an asynchronous kernel to initialize b
    myQueue.submit([&](auto &cgh) {
        // The kernel write b, so get a write accessor on it
        auto B = b.get_access<access::write>(cgh);
        /* From the access pattern above, the SYCL runtime detect
           this command_group is independant from the first one
           and can be scheduled independently */

        // Enqueue a parallel kernel on a N*M 2D iteration space
        cgh.parallel_for<class init_b>({ N, M },
                                       [=] (auto index) {
                                         B[index] = index[0]*2014 + index[1]*42;
                                       });
      });
    // Launch an asynchronous kernel to compute matrix addition c = a + b
    myQueue.submit([&](auto &cgh) {
        // In the kernel a and b are read, but c is written
        auto A = a.get_access<access::read>(cgh);
        auto B = b.get_access<access::read>(cgh);
        auto C = c.get_access<access::write>(cgh);
        // From these accessors, the SYCL runtime will ensure that when
        // this kernel is run, the kernels computing a and b completed

        // Enqueue a parallel kernel on a N*M 2D iteration space
        cgh.parallel_for<class matrix_add>({ N, M },
                                           [=] (auto index) {
                                             C[index] = A[index] + B[index];
                                           });
      });
    /* Request an access to read c from the host-side. The SYCL runtime
       ensures that c is ready when the accessor is returned */
    auto C = c.get_access<access::read, access::host_buffer>();
    std::cout << std::endl << "Result:" << std::endl;
    for(size_t i = 0; i < N; i++)
      for(size_t j = 0; j < M; j++)
        // Compare the result to the analytic value
        if (C[i][j] != i*(2 + 2014) + j*(1 + 42)) {
          std::cout << "Wrong value " << C[i][j] << " on element "
                    << i << ' ' << j << std::endl;
          exit(-1);
        }
  } /* End scope of myQueue, this wait for any remaining operations on the
       queue to complete */
  std::cout << "Good computation!" << std::endl;
  return 0;
}
```

Pipes in OpenCL 2.*x*

- Simple FIFO objects
- Useful to create dataflow architectures between kernels without host
- Created on the host with some message size + object number
- `read()`/`write()` functions
- Same behaviour/guaranty as a memory buffer
  - For portability because no hardware FIFO mandatory in OpenCL 2.*x*
  - Can be implemented with a memory buffer
- Non blocking because no independent-forward-progress guaranty in execution model yet
  - No guaranty that a producer can run concurrently with a consumer
  - No guaranty between different work-items when blocking

Pipes on FPGA

- The actual motivation for pipes in OpenCL standard!
- External memory access main cause for power consumption... ☹
- Real FIFO are easy to implement in hardware
  - Simple bus for 1-element FIFO
  - Latches or memory when more elements
- Very energy efficient
- Possible to have full dataflow applications without host control
- ⇝ FPGA vendors provide OpenCL extensions for pipe with stronger guarantees
  - Blocking pipes ⇝ simpler applications
  - Static size possible ⇝ direct synthesis
  - Independent work-groups and kernels for producers/consumers connected with pipes

⇝ Xilinx evaluates pipe extensions for SYCL too

Producer/consumer with blocking pipe

```
#include <CL/sycl.hpp>
#include <iostream>
#include <iterator>

constexpr size_t N = 3;
using Vector = float[N];

int main() {
  Vector va = { 1, 2, 3 };
  Vector vb = { 5, 6, 8 };
  Vector vc;

  {
    // Create buffers from a & b vectors
    cl::sycl::buffer<float> ba { std::begin(va), std::end(va) };
    cl::sycl::buffer<float> bb { std::begin(vb), std::end(vb) };

    // A buffer of N float using the storage of vc
    cl::sycl::buffer<float> bc { vc, N };

    // A pipe of 2 float elements
    cl::sycl::pipe<float> p { 2 };

    // Create a queue to launch the kernels
    cl::sycl::queue q;

    // Launch the producer to stream A to the pipe
    q.submit([&](cl::sycl::handler &cgh) {
      // Get write access to the pipe
      auto kp = p.get_access<cl::sycl::access::write,
                             cl::sycl::access::blocking_pipe>(cgh);
      // Get read access to the data
      auto ka = ba.get_access<cl::sycl::access::read>(cgh);

      cgh.single_task<class producer>([=] {
          for (int i = 0; i != N; i++)
            kp << ka[i];
        });
      });

    // Launch the consumer that adds the pipe stream with B to C
    q.submit([&](cl::sycl::handler &cgh) {
      // Get read access to the pipe
      auto kp = p.get_access<cl::sycl::access::read,
                             cl::sycl::access::blocking_pipe>(cgh);

      // Get access to the input/output buffers
      auto kb = bb.get_access<cl::sycl::access::read>(cgh);
      auto kc = bc.get_access<cl::sycl::access::write>(cgh);

      cgh.single_task<class consumer>([=] {
          for (int i = 0; i != N; i++)
            kc[i] = kp.read() + kb[i];
        });
      });
  } /*< End scope for the queue and the buffers:
       wait for completion q completion & bc copied back to v */

  std::cout << std::endl << "Result:" << std::endl;
  for(auto e : vc)
    std::cout << e << "_";
  std::cout << std::endl;
}
```

Non blocking pipe

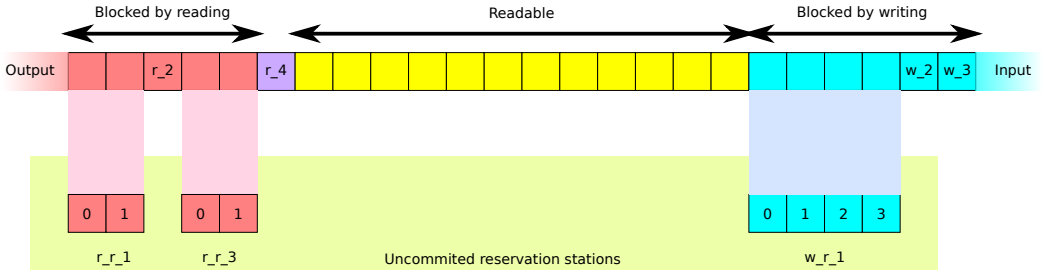
```
// Launch the producer to stream A to the pipe
q.submit([&](cl::sycl::handler &cgh) {
  // Get write access to the pipe
  auto p = P.get_access<cl::sycl::access::write>(cgh);
  // Get read access to the data
  auto ka = A.get_access<cl::sycl::access::read>(cgh);

  cgh.single_task<class producer>([=] {
      for (int i = 0; i != N; i++)
        // Try to write to the pipe up to success
        while (!(p << ka[i]))
          ;
    });
  });
```

Sequential pipe access

- FIFO: queue with serialized access
- How to implement simultaneous access by several work-item either at input or output?
- How to order access by work-items for deterministic execution?

⇝ Add concept of reservation station

Parallel pipe access with reservation



- Reservation station $\equiv$ array-view reserved in the pipe
- Allow ordered and parallel operation inside pipes
- Accessible up to commit operation
- Several reservation stations alive in parallel
- Can be mixed with on-going simple pipe access

Code example with reservation station

```
// Size of the buffers
constexpr size_t N = 200;
// Number of work-item per work-group
constexpr size_t WI = 20;
// The plumbing with some weird size prime to WI to exercise the system
cl::sycl::pipe<Type> pa { 2*WI + 7 };
// A buffer of N Type to get the result
cl::sycl::buffer<Type> c { N };
q.submit([&] (cl::sycl::handler &cgh) {
    // Get read access to the pipe
    auto apa = pa.get_access<cl::sycl::access::read,
                             cl::sycl::access::blocking_pipe>(cgh);
    // Get write access to the data
    auto ac = c.get_access<cl::sycl::access::write>(cgh);

    /* Create a kernel with WI work-items executed by work-groups of
       size WI, that is only 1 work-group of WI work-items */
    cgh.parallel_for_work_group<class consumer>(
      { WI, WI },
      [=] (auto group) {
        // Use a sequential loop in the work-group to stream chunks in order
        for (int start = 0; start != N; start += WI) {
          auto r = apa.reserve(WI);
          group.parallel_for_work_item([=] (cl::sycl::item<> i) {
              ac[start + i[0]] = r[i[0]];
            });
          // Here the reservation object goes out of scope: commit
        }
      });
  });
```

Exascale-ready

- Use your own C++ compiler
  - Only kernel outlining needs SYCL compiler
- SYCL as plain single-source C++ can address most of the hierarchy levels
  - MPI
  - OpenMP
  - C++-based PGAS (Partitioned Global Address Space) DSEL (Domain-Specific embedded Language, such as Coarray C++...)
  - Remote accelerators in clusters
  - Use SYCL buffer allocator for
    - RDMA
    - Out-of-core, mapping to a file
    - PiM (Processor in Memory)
    - ...

Using SYCL-like models in other areas

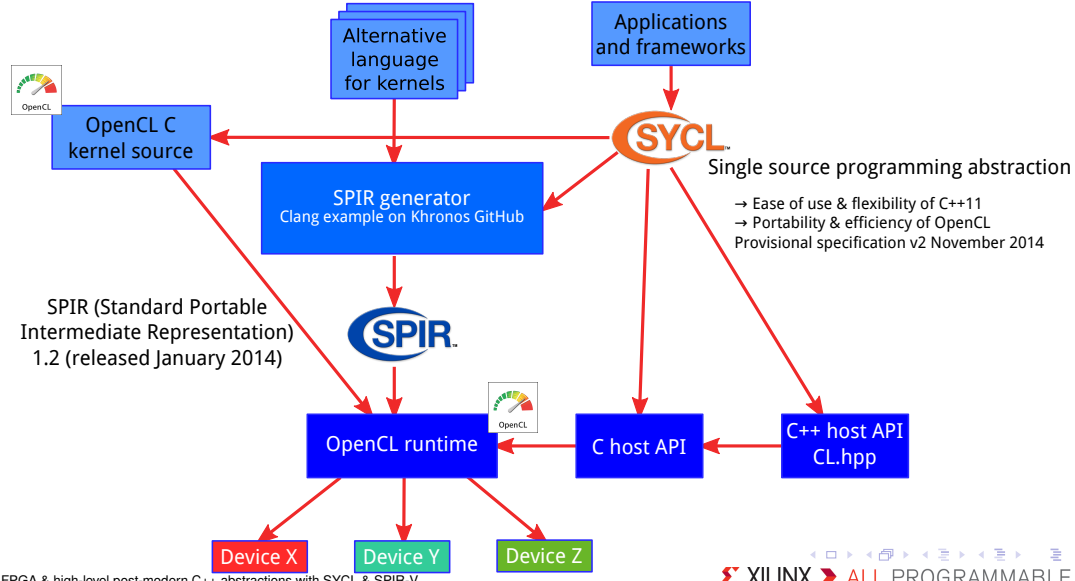
- SYCL ≡ generic heterogeneous computing model beyond OpenCL
  - `device` abstracts the accelerators
  - `queue` allows to launch tasks with computations overlapping communications and pipelining
  - `parallel_for<>` for // computations
  - `accessor` defines the way we access data
  - `buffer` to chose where to store data
  - `allocator` for defining how data are allocated/backed and how pointers work
- Example in PiM (Processor-in-Memory) world
  - Use `queue` to run on some PiM chips
  - Use `allocator` to distribute data structures or to allocate `buffer` in special memory (memory page, chip...)
  - Use `accessor` to use alternative data access (split address from computation, streaming only, PGAS...)
  - Use `pointer_trait` to use specific way to interact with memory such as bank/transposition or relocation

SYCL C++ for FPGA

- Xilinx FPGA
  - Clocks
  - AXI ports
  - AXI stream ports
  - Interrupt ports
  - I/O devices (Ethernet, Interlaken, ADC, DAC, video...)
  - Reset
  - Dynamic Voltage and Frequency scaling (DVFS)
  - Dynamic reconfiguration
  - Kernel scheduling
- Use native kernels to access specific I/O & IP
  - Single source C++  hidden in “normal” class interface

- Add location/placement in device & sub-device selectors
- Use C++11 allocators to select memory type & location
- Use accessors to define read/write/bus (AXI4 master/slave/...)/pipe/linear/... data access
- Use SystemC-like data types for user-defined size & precision
- C++: normal API to control run-time & OS
- Tool metadata can be moved optionally from XML/TCL/JSON/... into C++ classes
  - Metaprogramming HLS ☺

SYCL in OpenCL ecosystem



Parallel STL towards C++17 proposal

- Current Parallel STL from C++17 proposal N4507 (2015/05/05)
  `http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2015/n4507.pdf`

```
// Current C++11: standard sequential sort
std::sort(vec.begin(), vec.end());
// C++17: permitting parallel execution and vectorization as well
sort(std::experimental::parallel::par_vec, vec.begin(), vec.end());
```

- Easy to implement in SYCL
  - Could even be extended to give a kernel name (profile, debug...):
  - Load balancing between CPU and accelerator

```
sycl_policy<class kernelName1> pol;
sort(pol, begin(vec), end(vec));

sycl_policy<class kernelName2> pol2;
// But SYCL allows OpenCL intrinsics in the operation too
for_each(pol2, vec.begin(), vec.end(),
         [](float & ans) { ans += cl::sycl::sin(ans); });
```

Open Source `https://github.com/KhronosGroup/SyclParallelSTL`

Outline

Known implementations of SYCL

- ComputeCPP by Codeplay `https://www.codeplay.com/products/computecpp`
  - Most advanced SYCL 1.2 implementation
  - Outlining compiler generating SPIR
  - Run on any GPU and CPU
- sycl-gtx `https://github.com/ProGTX/sycl-gtx`
  - Open source
  - No (outlining) compiler ⇝ use some macros with different syntax
- triSYCL `https://github.com/amd/triSYCL`
  - Open Source
  - Some extensions (Xilinx blocking pipes)
  - No (outlining) compiler ⇝ no device support yet

Standard is still moving & no full implementation yet

triSYCL

- Open Source implementation using templated C++1z classes
  - On-going implementation started at AMD and now lead by Xilinx
  - `https://github.com/amd/triSYCL`
  - 7 contributors
- Used by Khronos committee to define the standard
  - Languages are now too complex to be defined without implementation
- Pure C++ implementation & CPU-only implementation for now
  - Use OpenMP for computation + `std::thread` for task graph
  - Rely on STL & Boost for zen style
  - Quite useful for debugging
  - CPU emulation for free
- Looking for some interns ☺ to add outlining compiler to generate SPIR-V based on open source Clang/LLVM, etc.

Conclusion

- Modern FPGA are complex MP-SoC
- Full systems & HPC machines add another complexity level
- Modern C++ can be used for portable single-source DSEL targeting heterogeneous systems
- SYCL C++ Khronos standard provides seamless single source with OpenCL interoperability
  - Can be used to improve other higher-level frameworks
- SYCL $\equiv$ pure C++ $\rightsquigarrow$ integration with other C/C++ HPC frameworks: OpenCL, OpenMP, libraries (MPI, numerical), C++ DSeL (PGAS...)...
- SYCL interesting as co-design tool for architectural & programming model exploration (PiM, Near-Memory Computing, FPGA, various computing models...)
  - Inspirational to future OpenCL C++ kernel language
  - Built on top of open source projects & HLS tools
    - SPIR-V gives portable execution model
- Modern C++ is not just C program in `.cpp` file ☺ $\rightsquigarrow$ Invest in learning modern C++
  - But can also use SYCL in a old-C style programming... ☺