FPGA fabric is eating the world

The rise of the custom computing machines

From the eyes of Steve Casselman

What is the FABRIC?

- Fabric is the sum of all the hardware in a computing system
- In the beginning the Fabric was simple; an ALU and some controllers
- The Fabric grew, and there were different kinds of Fabric; vector machine, big iron, and finally clusters
- You can also think about the Fabric of a single device
- In the beginning devices were simple; an ALU and some controllers
- Then came Main Frame cores, Mini CPUs, Micro CPUs, then FPGAs and finally GPUs
- This talk is about the past, present and future of reconfigurable computers and the FPGA fabric on which they are based

We define reconfigurable computing as

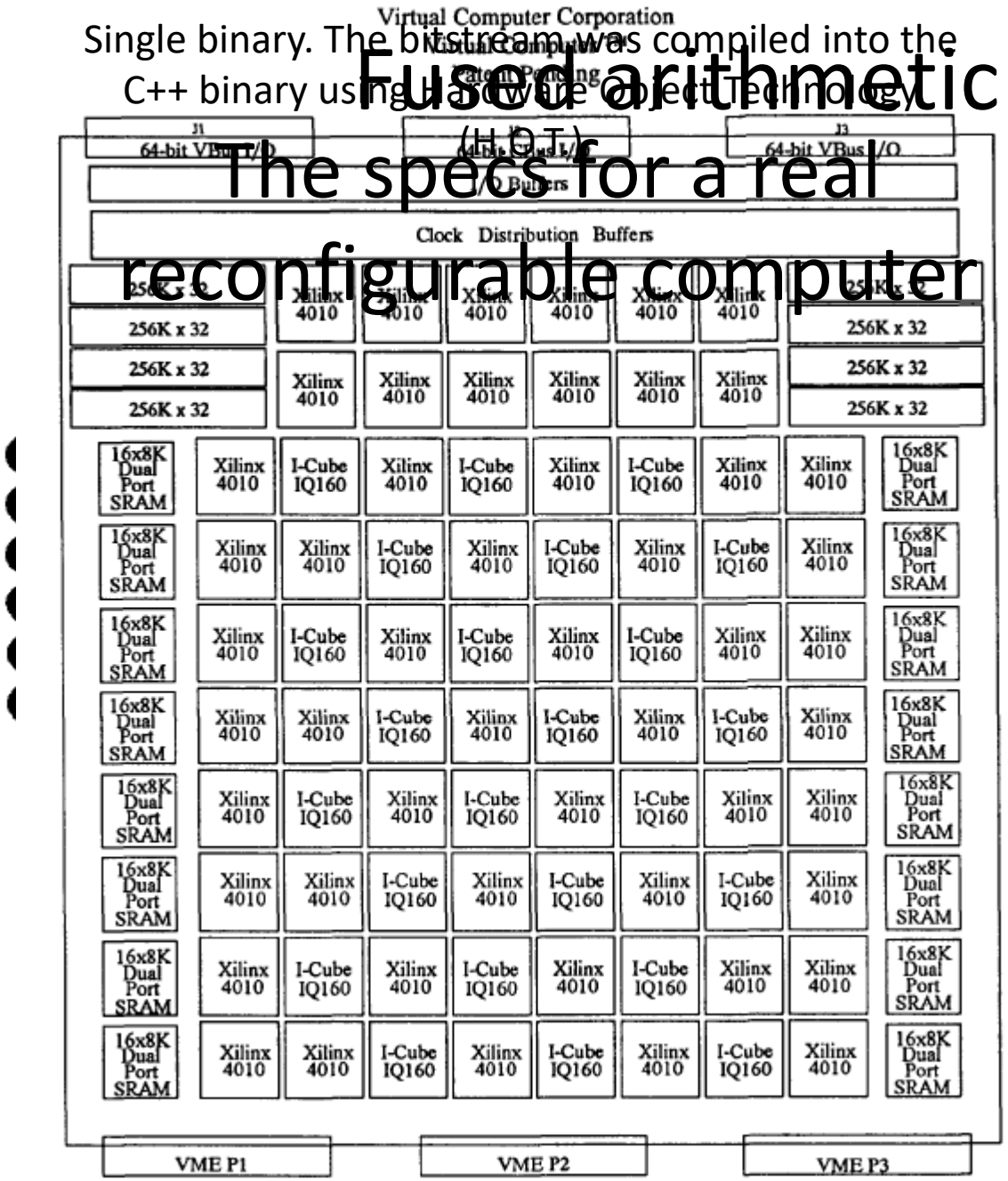
- taking a high-level language
- compiling it to an FPGA bitstream
- and running those bitstreams one after another

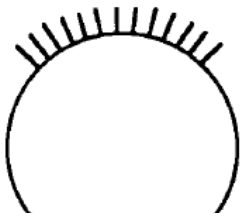
From my paper at the first FCCM in 1992 “Virtual Computing and The Virtual Computer”

Single binary. The bitstream was compiled into the C++ binary using Hardware Object Technology (HOT).

Fused arithmetic

The specs for a real reconfigurable computer

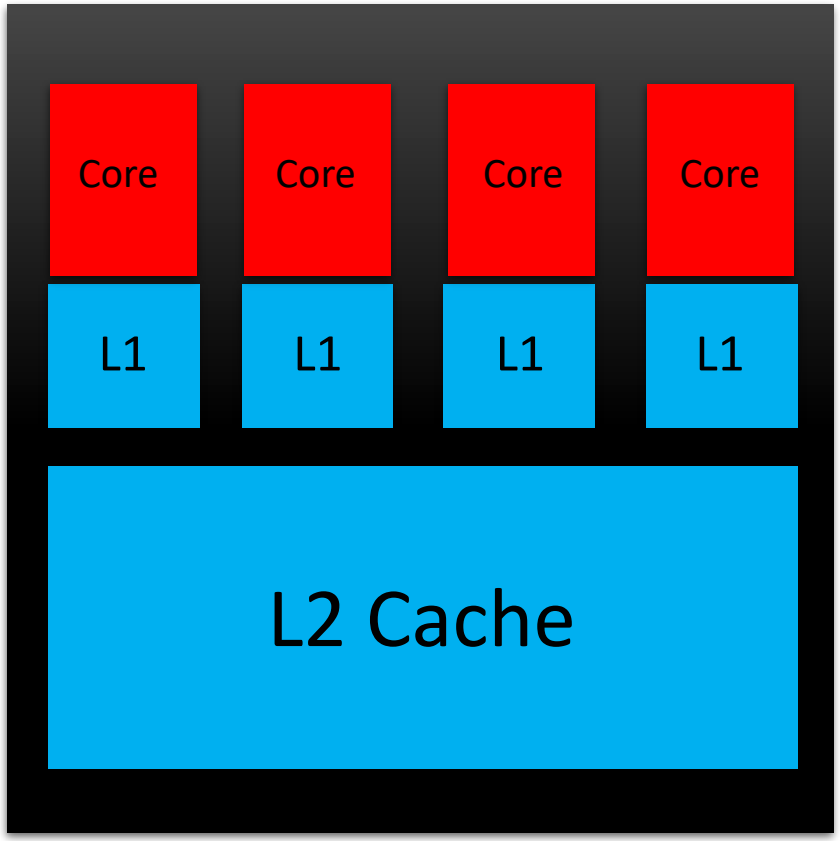




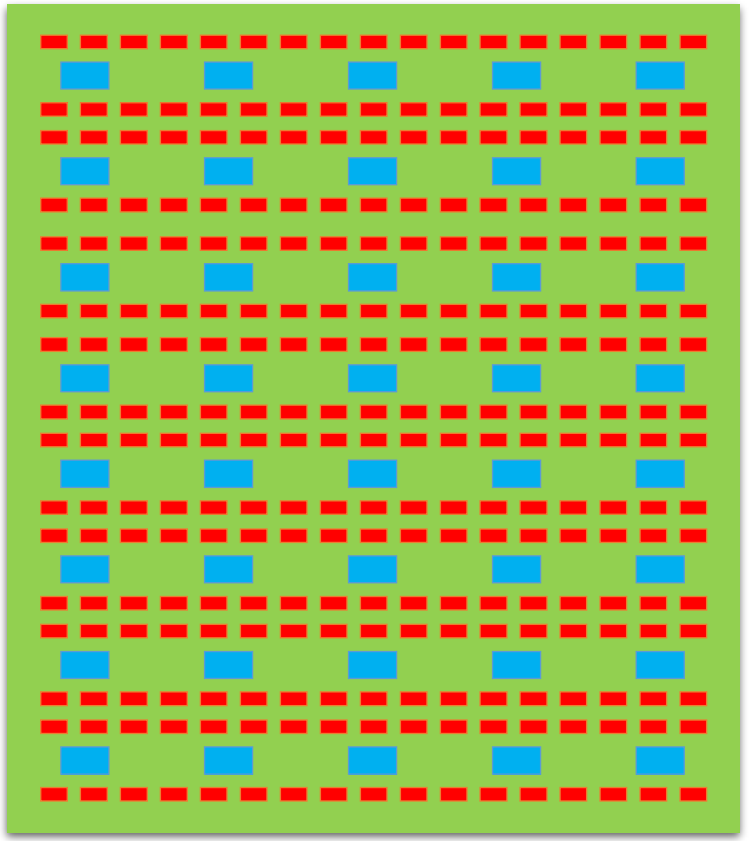
mory
ficients
in to any other pin)

Why are FPGAs good for computing?

“The UCSD Center for Dark Silicon was among the first to demonstrate the existence of a utilization wall which says that with the progression of Moore's Law, the percentage of a chip that we can actively use within a chip's power budget is dropping *exponentially*! The remaining silicon that must be left unpowered is now referred to as Dark Silicon.”
This is also known as the breakdown of Dennard scaling!



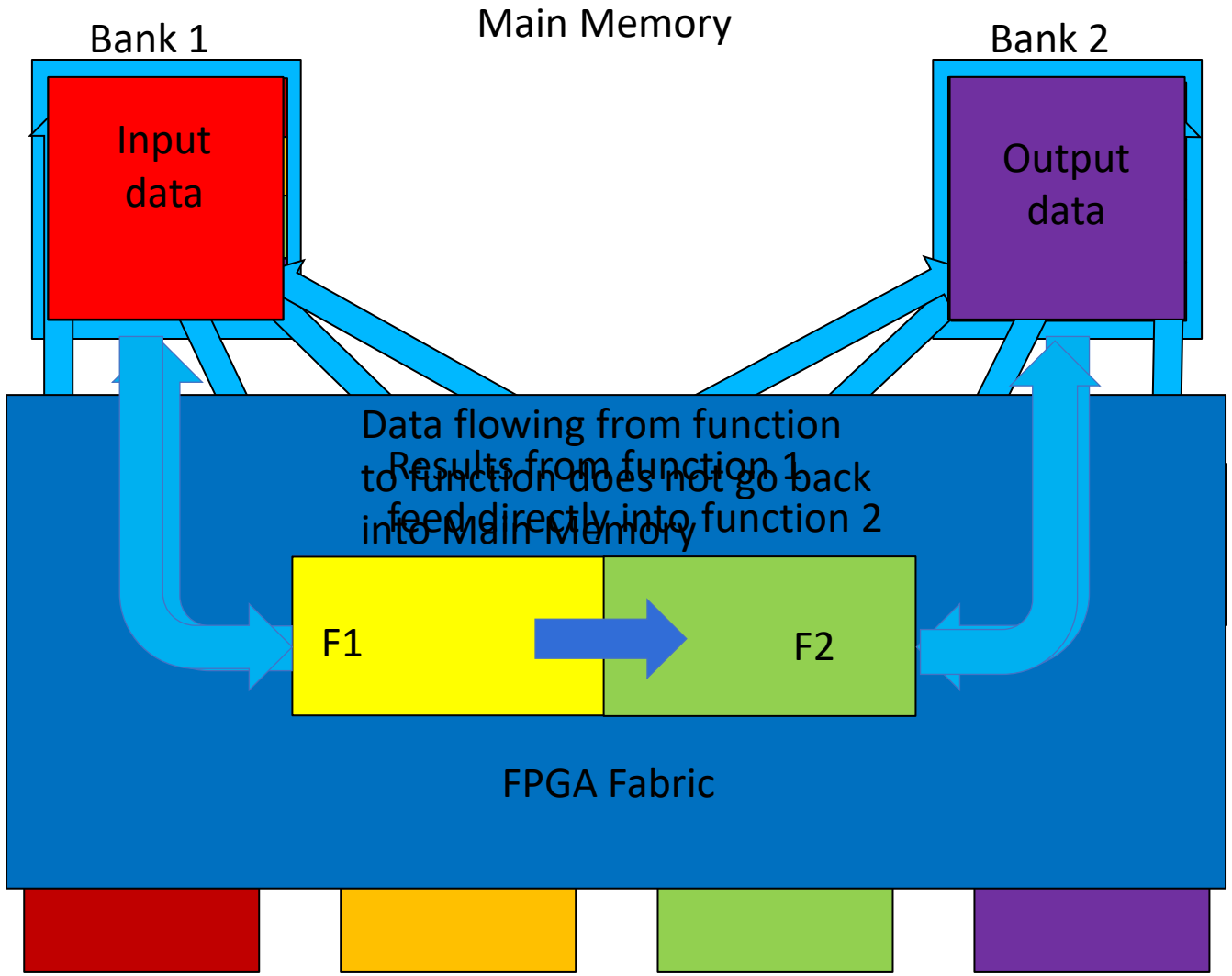
High speed CPU (or GPU) cores get very hot. So hot they fail



Compute power is spread out and performance comes from pipelining. The logic is in red and memory in blue

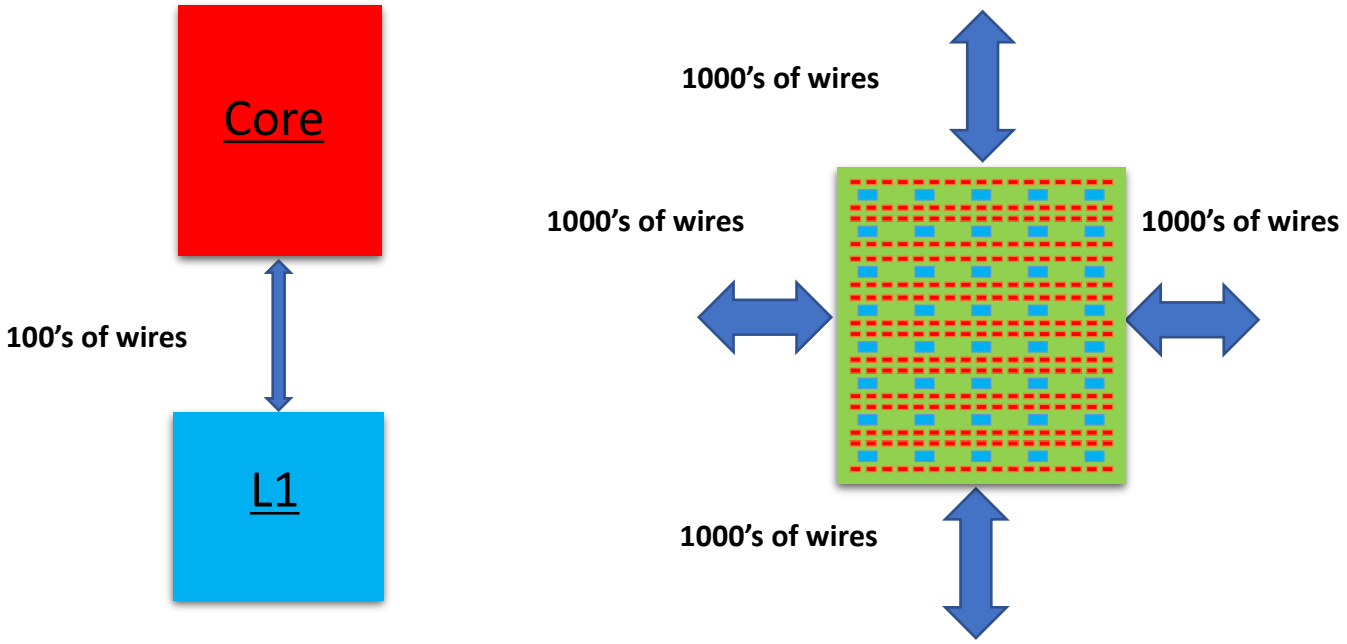
Each core in a multicore processor system shares main memory with the other cores. Lots of data collisions and congestion.

- Results can be used directly by the next function without going back to memory. Result reuse lowers memory access and therefor overall power usage in regards to TCO.



Rent's Rule

Rent's rule describes the relationship between the amount of logic in a partition and the amount of communication into that partition. FPGAs are architected based on Rent's rule and CPUs and GPUs are not. The logic cores of CPUs and GPUs are connected to caches through which the data must pass.



FPGAs, on the other hand, have 1000's of wires coming into a logic partition from all directions. Data flow in FPGAs is managed through 100's to 1000's of custom connected multi-ported memories instead of a hierarchical memory system based on different levels of cache.

The 6 waves of reconfigurable computing

- Invention of FPGA. (event)
  - Ross Freeman.

Ross Freeman started it all

- In 1984 Ross Freeman and his band of engineers created the first commercially successful FPGA
- The device used memories, registers and pass transistors to create a homogenous array of lookup table (LUT) logic and changeable routing
- The device was based on SRAM and so could be reconfigured on demand
- Device support for reconfigurable computing was not there in the beginning.
  - A PAL was needed next to the device to make it into a reconfigurable computer
- That’s what I did

The 6 waves of reconfigurable computing

- Invention of FPGA. (event)
  - Ross Freeman.
- Invention of Reconfigurable Computing 1st company VCC (pre wave stealth)

Steve Casselman's introduction to FPGAs

- In 1986 someone came into the EDA lab, spotted me and said "Casselman you like weird stuff, come out and talk to this new vendor with me"
- The new vendor was Monolithic Memories Inc, which was a second source for Xilinx
- The new part was called a Logic Cell Array (LCA)
- This was before they had schematic capture for design entry
- I knew right away that the LCA was a new kind of processor with a weird programming model
- I was sure it could be programmed because "Anything you can do in hardware you can do in software and vice versa"

What happened when I started in 1986

- Challenger
- Halley's Comet
- Microsoft IPO
- Chernobyl
- Iran-Contra
- Born that year
  - Lady Gaga
  - Lindsay Lohan

Before the first wave

PRINCIPAL INVESTIGATORS (NAME AND TITLE)
Steven M. Casselman

TITLE OF PROJECT
A Fully Programmable Reconfigurable Hardware Architecture Supercomputer.

TOPIC TITLE	TOPIC NUMBER
NEW COMPUTING DEVICES	15 D

TECHNICAL ABSTRACT (LIMIT TO 200 WORDS)

The research proposed is investigation of a new approach into the area of supercomputing. With the advent of the programmable gate-array, [1] the possibility of mapping a software program directly into a large number of such devices implies a significant advance in the area of supercomputing. This ability to repeatably map software directly into a fully reconfigurable hardware architecture will minimize many of the problems facing conventional and parallel supercomputing such as memory fetch, microcode memory fetch, and sequencer decoding delay.

The research to be done will be two-fold:

1) Study the topology of the interconnection of arrays to find a way to allow a continuous plane of arrays to be created.

2) Write a compiler that will map a source code file into the proper binary format needed by the arrays.

1987 SBIR

KEY WORDS TO IDENTIFY RESEARCH OR TECHNOLOGY (8 MAXIMUM)
Supercomputer, Reconfigurable Hardware.

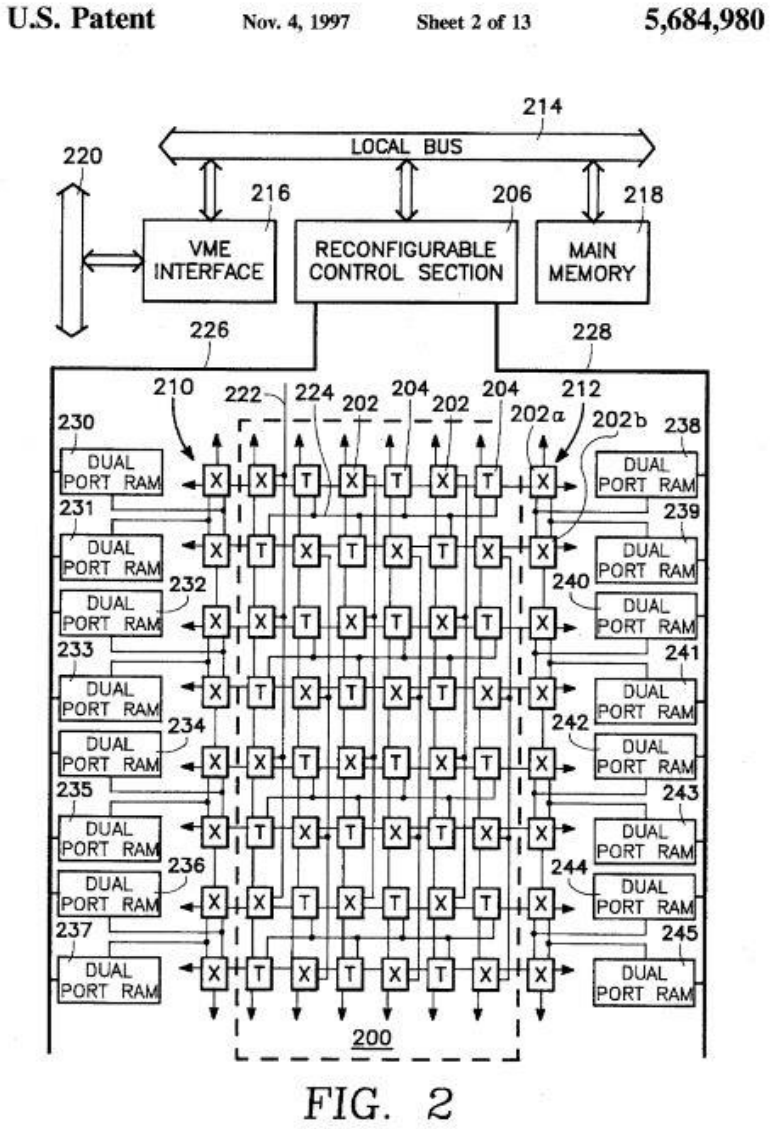
POTENTIAL COMMERCIAL APPLICATIONS OF THE RESEARCH

Commercial applications range from use in highly recursive mathematical and physical problems to high speed simulation of large systems.

The 6 waves of reconfigurable computing

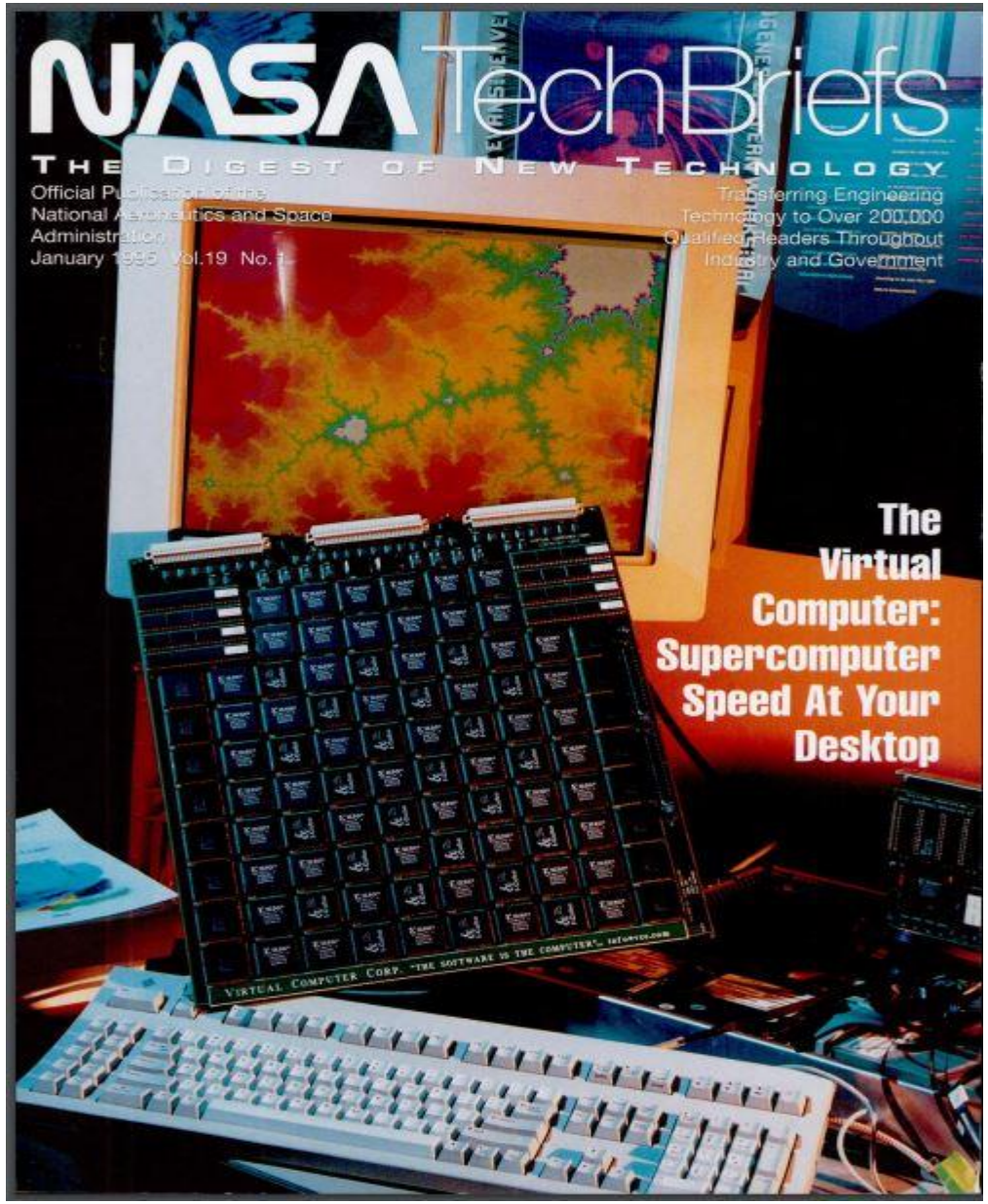
- Invention of FPGA. (event)
  - Ross Freeman.
- Invention of Reconfigurable Computing 1st company VCC (pre wave stealth)
- The first wave, NASA Technology Briefs, EETimes and a couple of conferences

First wave



My first patent was filed in 1992 granted in 1997

We won the first SBIR of the year



Virtual Computer Corporation (VCC), Reseda, CA, developed its award-winning Virtual Computer™ for the US Naval Surface Warfare Department in 1991. The device is one of a new class of computing machine called reconfigurable hardware. This class employs massively reconfigurable, or programmable, logic, blurring the line between hardware and software.

The technology solves the problem of Amdahl's law, which limits performance improvements on hardware scale-ups:

For more information about the 1995 SBIR Technology of the Year competition, contact Wayne Pierce, Technology Utilization Foundation, 41 East 42nd St., Suite 921, New York, NY 10017. Tel: 212-490-3999; Fax: 212-986-7864.

First SBIR technology of the year, 1995

The 6 waves of reconfigurable computing

- Invention of FPGA. (event)
  - Ross Freeman.
- Invention of Reconfigurable Computing 1st company VCC (pre wave stealth)
- The first wave, NASA Technology Briefs, EETimes and a couple of conferences
- Second Wave Many conferences, 2nd wave of small businesses, early press

Darpa said “We will bring you the future”

We made a deal with the distributor to source all the components for the board

We then packaged the board with our software, and the distributor stocked and sold all systems

In a Scientific American article DARPA promised to invent the future.

In the same issue we offered the future for sale

High level programming languages come online

- Handel C
  - Ian Page
- Napa Compiler
  - Maya Gokhale, Jeff Arnold
- JBits
  - Steve Guccione
  - One of the most important projects in reconfigurable computing history
  - JBits generates a bitstream, deterministically, in less than a second

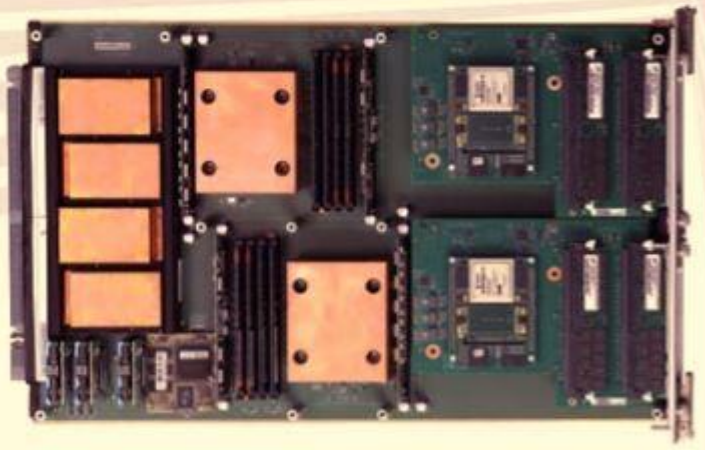
The 6 waves of reconfigurable computing

- Invention of FPGA. (event)
  - Ross Freeman.
- Invention of Reconfigurable Computing 1st company VCC (pre wave stealth)
- The first wave, NASA Technology Briefs, EETimes and a couple of conferences
- Second Wave Many conferences, 2nd wave of small businesses, early press
- Third wave – real money: Comm processors – end of 3rd wave small companies get bought up, AI inference works best on FPGA

FPGAs deployed in a supercomputer

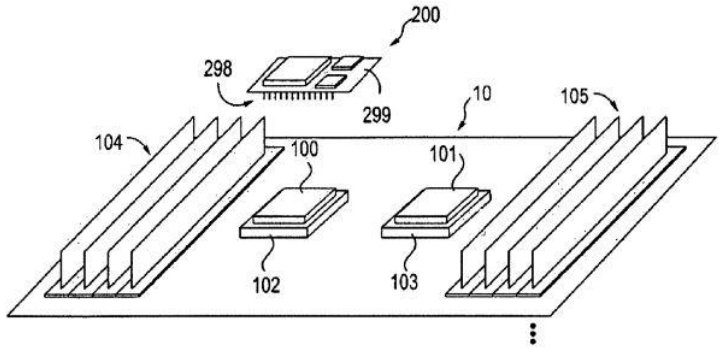
The Cray XR1 Reconfigurable Processing Blade, compatible with existing Cray XT3™ and Cray XT4™ systems as well as new Cray XT5$_h$™ systems, offers users orders of magnitude speedup on select applications as well as large potential savings in cooling and space. Building on the established track record of the Cray XT™ product line and the reconfigurable computing capability in the Cray XD1™ system, the Cray XR1 reconfigurable processing blade is the first product on the market capable of massively parallel reconfigurable computing.

Cray XR1 Reconfigurable Processing Blade



High Bandwidth, Direct Connect Architecture

A Cray XR1 reconfigurable blade has two nodes, consisting of a single AMD Opteron™ processor tightly coupled with two DRC Computer's reconfigurable processing units (RPUs). This connection is made directly with HyperTransport™, which ensures that RPUs are tightly coupled with AMD Opterons, delivering low-latency and high-bandwidth communication between the processing elements.



The FPGA in the processor socket patent was filed in 2007

OEMed by Cray
Bought by the Australian and New Zealand secret services.

More high-level programming languages come online

- AutoESL
  - Jason Cong
  - Becomes the basis for Xilinx HLS
- Catapult C
  - Mentor
- Impulse C
  - Dave Pellerin
  - I used this to get 80x on one project
  - One part of the puzzle that convinced Microsoft to adopt FPGAs

Small companies that were bought or acquired

- Molex buys both Bittware and Nallatech
- Micron buys both Pico and Convey and
- DRC gets acquired by its largest customer

The 6 waves of reconfigurable computing

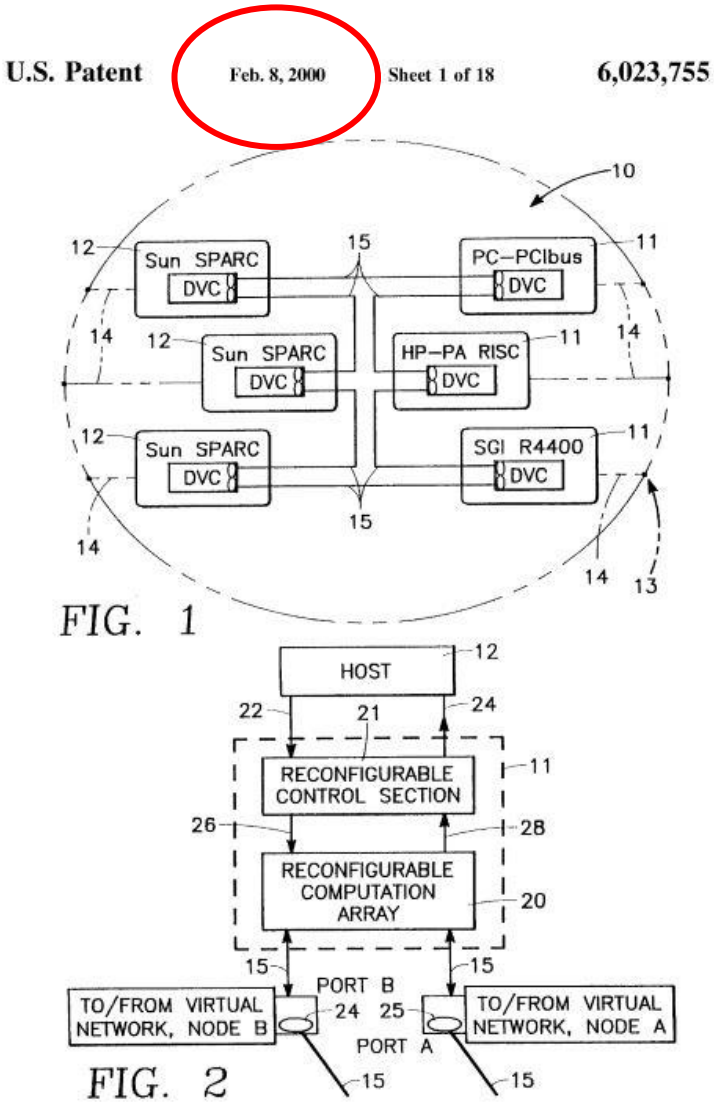
- Invention of FPGA. (event)
  - Ross Freeman.
- Invention of Reconfigurable Computing 1st company VCC (pre wave stealth)
- The first wave, NASA Technology Briefs, EETimes and a couple of conferences
- Second Wave Many conferences, 2nd wave of small businesses, early press
- Third wave – real money: Comm processors – end of 3rd wave small companies get bought up, AI inference works best on FPGA
- Forth wave – Today: big company buy in, Super 7, Azure, AWS 4th generation of small businesses appear

Distributed Virtual Computer (DVC)

The DVC allowed you to build system of directly connected FPGAs

Round trip latency was sub 2 microseconds a world record at the time.

Microsoft now uses this in all their new Azure Data Center Clusters



Combine FPGA + CPU

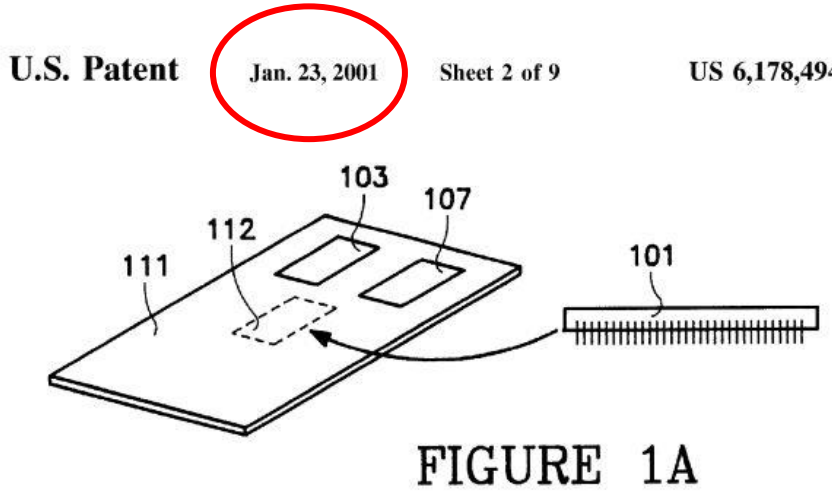


FIGURE 1A

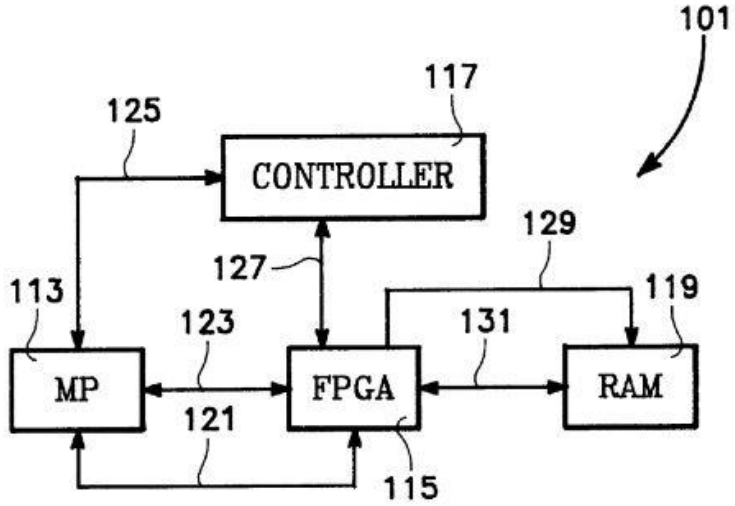


FIGURE 1B

This is Intel's and AMD's current plan

The 6 waves of reconfigurable computing

- Invention of FPGA. (event)
  - Ross Freeman.
- Invention of Reconfigurable Computing 1st company VCC (pre wave stealth)
- The first wave, NASA Technology Briefs, EETimes and a couple of conferences
- Second Wave Many conferences, 2nd wave of small businesses, early press
- Third wave – real money: Netezza, Comm processors – end of 3rd wave small companies get bought up, AI inference works best on FPGA
- Forth wave – Today: big company buy in, Super 7, Azure, AWS 4th generation of small businesses appear
- Fifth wave – total acceptance: FPGAs account for 20% of silicon in datacenter

The first 4 hits for the search "FPGA in the data center"

About 17,200,000 results (0.42 seconds)

www.nextplatform.com › Compute

The Inevitability Of FPGAs In The Datacenter - The Next Platform

Jan 14, 2020 — At some point, and the **FPGA** will probably usher this era along, we will go back to calling it data processing. Bringing **FPGAs** into the **datacenter** ...

www.intel.com › products › programmable › overview

Acceleration in the Data Center - Intel® FPGA

Unleash Your **Data Center**. For Intel® Xeon® CPU with **FPGAs**. Overview; Applications; Videos.

www.intel.com › content › www › programmable › fpg...

Intel FPGA Acceleration in the Data Center

... performance while minimizing power consumption in your **data center**? Learn how the Acceleration Stack for Intel® Xeon® CPUs with **FPGAs** and ...

www.xilinx.com › applications › data-center

Data Center Acceleration - Xilinx

Adaptable Acceleration for the Modern **Data Center**. Advances in artificial intelligence, increasingly complex workloads, and an explosion of unstructured data ...

More search results from page 1

www.rambus.com › Blogs › Smart Data Acceleration

The role of FPGA acceleration in the data center and beyond ...

Sep 20, 2016 — Gupta, the general manager of Xeon+**FPGA** products in Intel's **data center** group, said **FPGAs** can increase the performance of applications such ...

FPGA Acceleration Platform in a Data Center for a ... - arXiv

The field-programmable gate array (**FPGA**) is an ideal choice for maintaining the same infrastructure and provides customized computing architectures for different.

by X Yu · 2019 · Cited by 5 · Related articles

FPGAs in Data Centers - ACM Queue

Jun 5, 2018 — It is in this context that **FPGAs** have attracted the attention of system architects and have started to appear in commercial cloud platforms. An ...

Data Centers Get a Performance Boost from FPGAs

Aug 15, 2019 — HPE's Bill Mannel, explores how as big data continues to explode, **data centers** are benefitting from a relatively new type of offload accelerator: ...

Network-attached FPGAs for data center applications - IEEE ...

Abstract: **FPGAs** (Field Programmable Gate Arrays) are making their way into **data centers** (DC). They are used as accelerators to boost the compute power of ...

by J Weerasinghe · 2016 · Cited by 48 · Related articles

More ways to program hardware

- C/C++
- OpenCL
- OpenMP
- RapidWright
  - RapidWright.io is a Xilinx open-source project
  - Like JBits, you have access to the Basic Element (BEL) level
  - You can stitch together precompiled operators and functions
  - In seconds!
  - There is a real possibility of having a Just In Time (JIT) compiler for hardware!

The 6 waves of reconfigurable computing

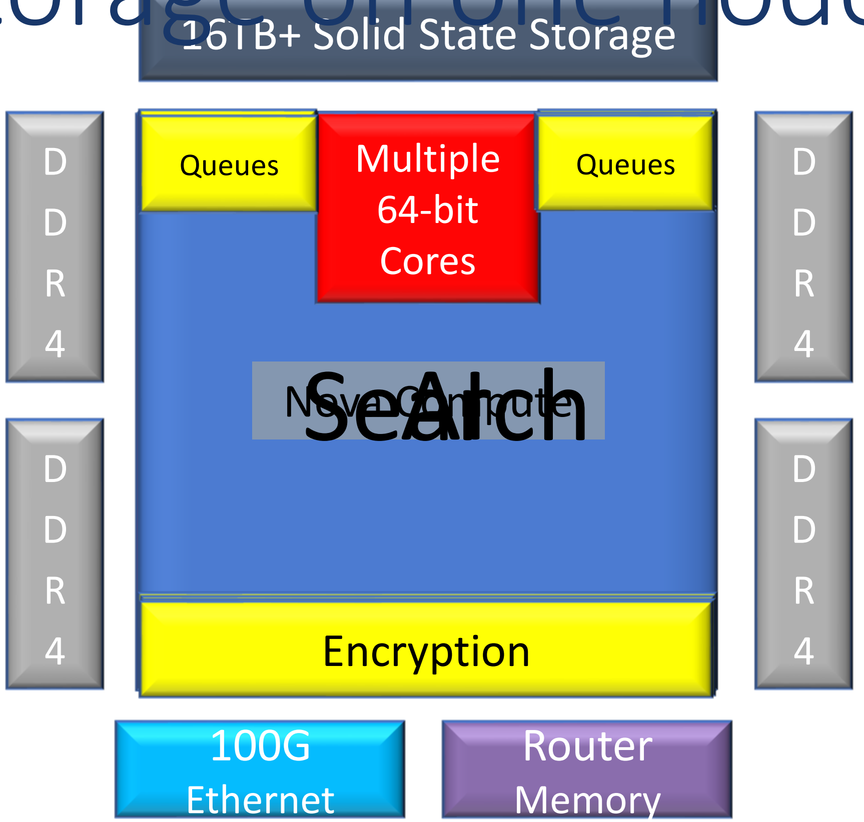
- Invention of FPGA. (event)
  - Ross Freeman.
- Invention of Reconfigurable Computing 1st company VCC (pre wave stealth)
- The first wave, NASA Technology Briefs, EETimes and a couple of conferences
- Second Wave Many conferences, 2nd wave of small businesses, early press
- Third wave – real money: Comm processors – end of 3rd wave small companies get bought up, AI inference works best on FPGA
- Forth wave – Today: big company buy in, Super 7, Azure, AWS 4th generation of small businesses appear
- Fifth wave – total acceptance: FPGAs account for 20% of silicon in datacenter
- Sixth wave – total dominance: wafer scale FPGA based systems account for 50+% of datacenter silicon

SDI OpenStack implementation:
Nova, Neutron & Swift
(Compute, Communication & Storage)

computation, communication and storage on one node

Swift storage functionally placed in hardware.

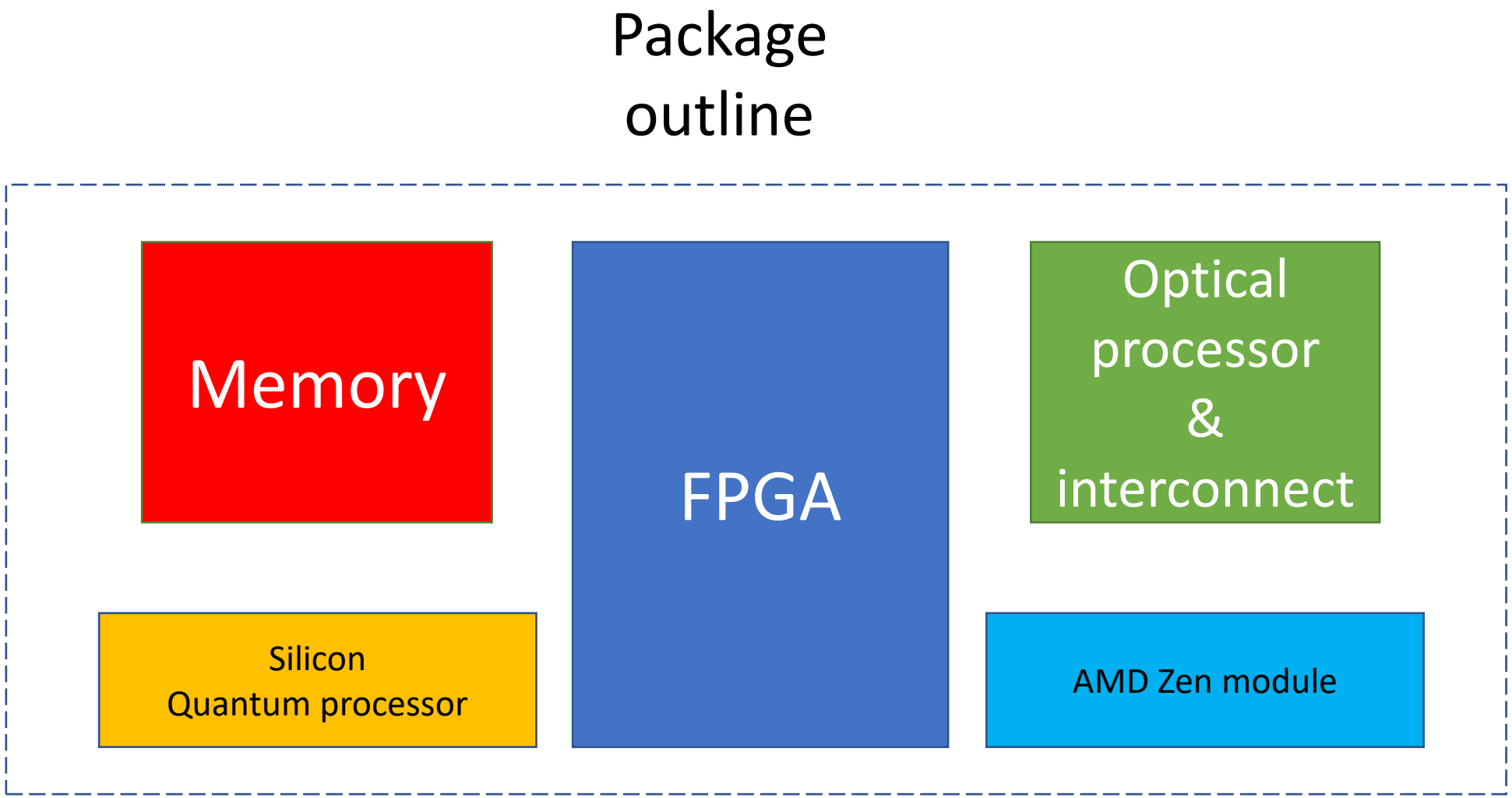
Neutron networking stack implemented directly in hardware.



Nova compute functions are mapped into CPU cores and FPGA fabric.

High random access HMC services: graph, pointer chasing and content addressable memory applications

Chiplet technology lets the fabric absorb everything



The future as seen by a visionary

Stacked wafers of FPGA fabric connected via fiber optics

Manufacturing flaws are put in a purge map

A vision from 1993 that gets better every day!

MRL COMPUTERS TURN ALGORITHMS INTO HARDWARE

When adding processors to massively parallel processing (MPP) systems, there is never a time when, by doubling the number of processors, you more than double the throughput of the system. That is loosely known as Amdahl's law or (if there is a 1:1 speedup) the law of perfect speedup.

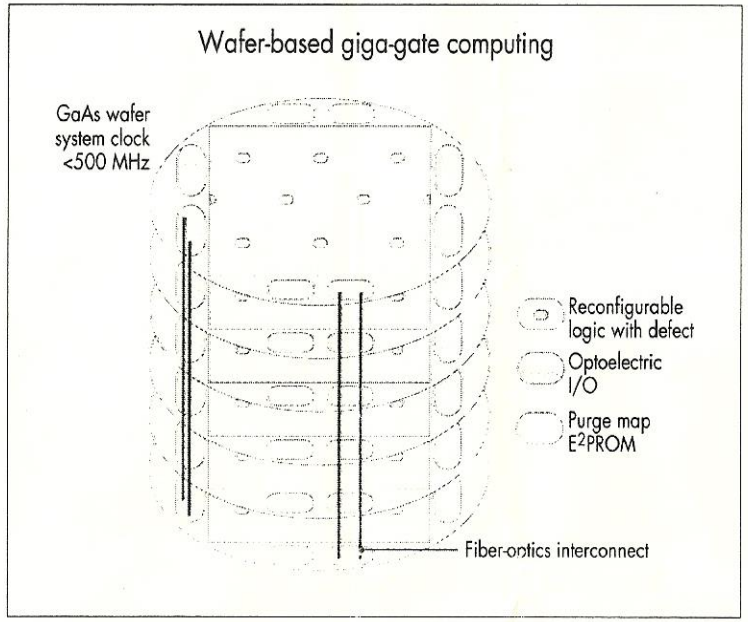
A computer architecture that could violate that law would be more than "perfect"—the computer-science equivalent of breaking the speed-of-light barrier in physics. Yet there is an architecture that does precisely that: massively reconfigurable logic (MRL).

An MRL computer can reconfigure its internal logic completely, in real-time, to implement an algorithm in hardware. It does so via field-programmable gate arrays. Downloading a file to the FPGAs rearranges the logic and routing resources inside to implement a hardware design.

The Supercomputing Research Center (SRC, Bowie, Md.) has already used the technique to build a machine that outperforms the Cray 2 by 330 times, operating on DNA-sequence comparisons.

Our version of an MRL computer, the Virtual Computer, is a single-board desktop machine with more than 500,000 gates of reconfigurable logic.

Since MRL systems use commercial, off-the-shelf parts, they are cheap, at $125,000. And with no moving parts, they can be offered with multiyear guarantees and reasonable repair cost estimates after that.



If a single transistor goes bad in a microprocessor, the whole chip is bad. In an MRL system, by contrast, a bad spot can be marked as not usable, much as in a hard disk's purge map. That will lead to the first efficient use of wafer-scale integration in which every wafer can be used.

Supercomputers in the year 2000 will be more open, more versatile and more reconfigurable than anyone can imagine at this time. Our vision for the future of computing is MRL-based Virtual Computers capable of 10^{15} operations/second at a cost of under $500,000.

—By Steven Casselman, president, Virtual Computer Corp. (Reseda, Calif.).

Every area of science must have a fundamental law

The fundamental law of FPGA fabrics is

"If a compute architecture is useful, it will be absorbed into the fabric"

Examples are:
Adders
Multipliers
Memories
High speed I/Os – PCIe, ethernet …
Processors
GPUs
Photonics, Optical computing
Quantum computing

FPGA Fabric is eating the world!

Thank you for your attention!