

The Modeling and Simulation Process for Large IBM Systems

José E. Moreira – IBM Research

Workshop on Modeling & Simulation of Systems and Applications
ModSim 2022 – August 10-12, 2022, Seattle, WA

Special thanks to Sameh Asaad and Charan Srivatsan

My career at IBM

- 1995-1999: Parallel job scheduling and management (lots of simulation)
- 1997-1999 : Java for numerical computing (little simulation)
- 2000-2005 : Blue Gene/L (lots of simulation)
- 2006-2007 : Commercial Scale Out (little simulation)
- 2008-2012 : POWER8 (lots of simulation)
- 2012-2017 : POWER9 (lots of simulation)
- 2018-2022 : POWER10 (lots of simulation)
- 2021- ... : Future POWER (lots of simulation)

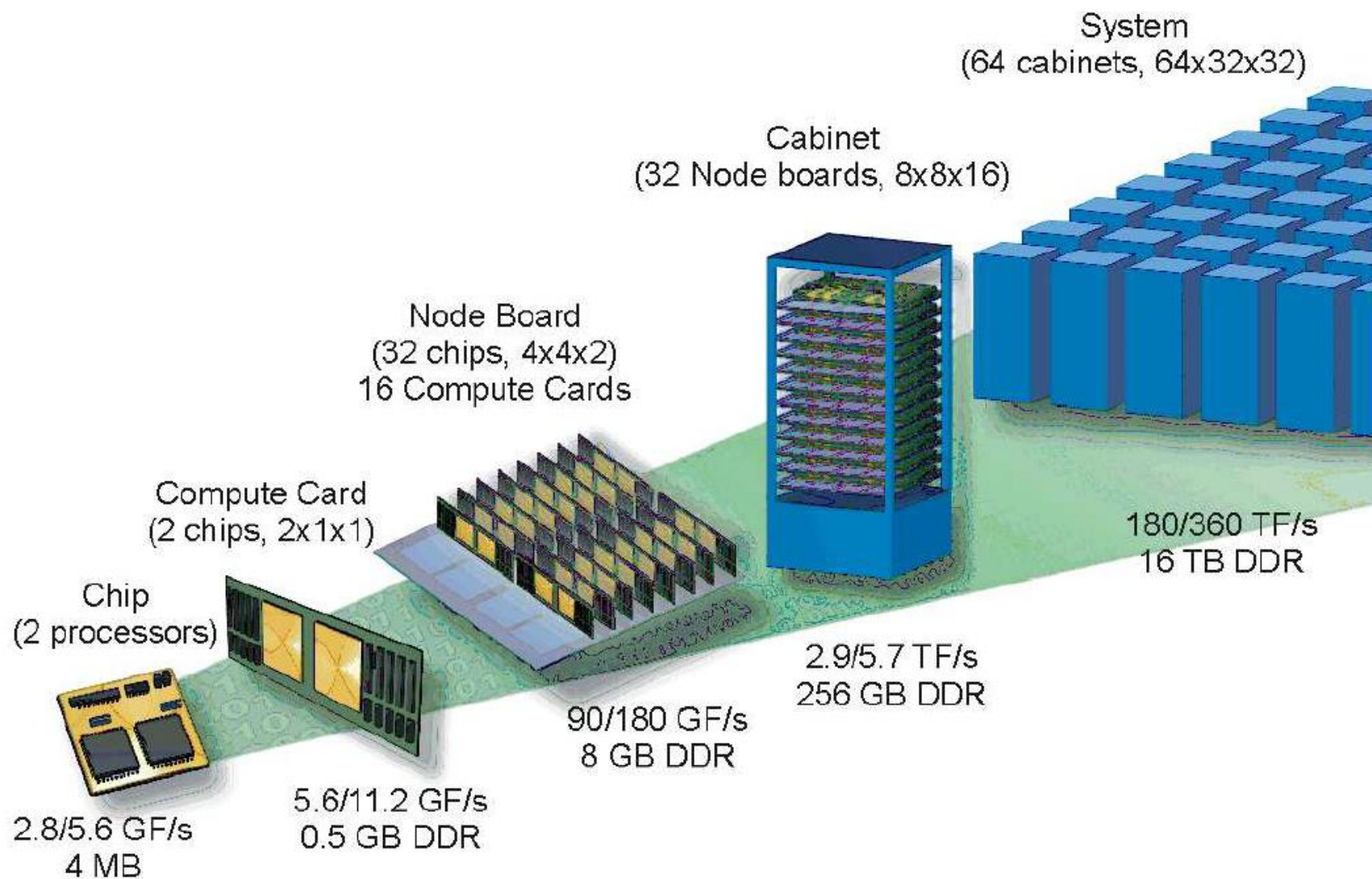
Motherhood and apple pie ...

- When developing new computing systems, whether Blue Gene or any of the POWER processors, simulation is essential
- Because so much must happen before any Silicon is available
- For the first three years of POWER10 project there was no Silicon
- And yet most important decisions were made during that period
- Simulation, at different levels, provides the input we need for design
- I will walk through *some* of the simulation and modeling methodology for two of our large systems that I worked on: Blue Gene/L and POWER10 – and there is *much that I will not cover*, such as power and thermal modeling

It starts with Mambo ...

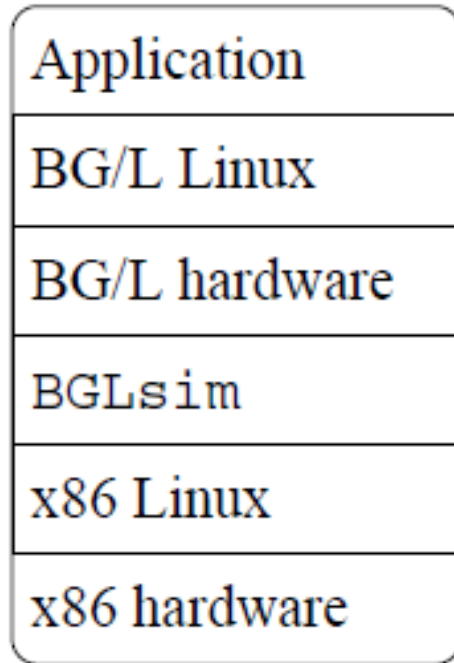
- Mambo (also called **systemsim**) is an ISA-level functional simulator
- A C program that runs on x86 and Power ISA systems
- It models a multi-core system with shared memory
- It serves multiple official purposes:
 - A reference ISA implementation, used for testing the design
 - A software development platform, for all new ISA features
 - A trace generation tool, even for existing systems
- It has been extended with performance models for specific projects:
 - Blue Gene (L/P/Q)
 - Transactional memory
 - Various cache models
 - IBM Exascale work
- It has been used to model large parallel systems

The Blue Gene/L supercomputer project (2000-2005)

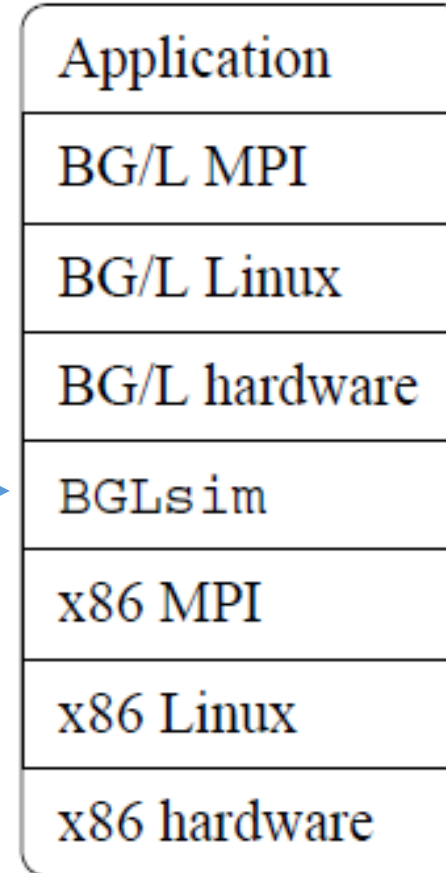


The Blue Gene/L system simulation stack

- Single-node



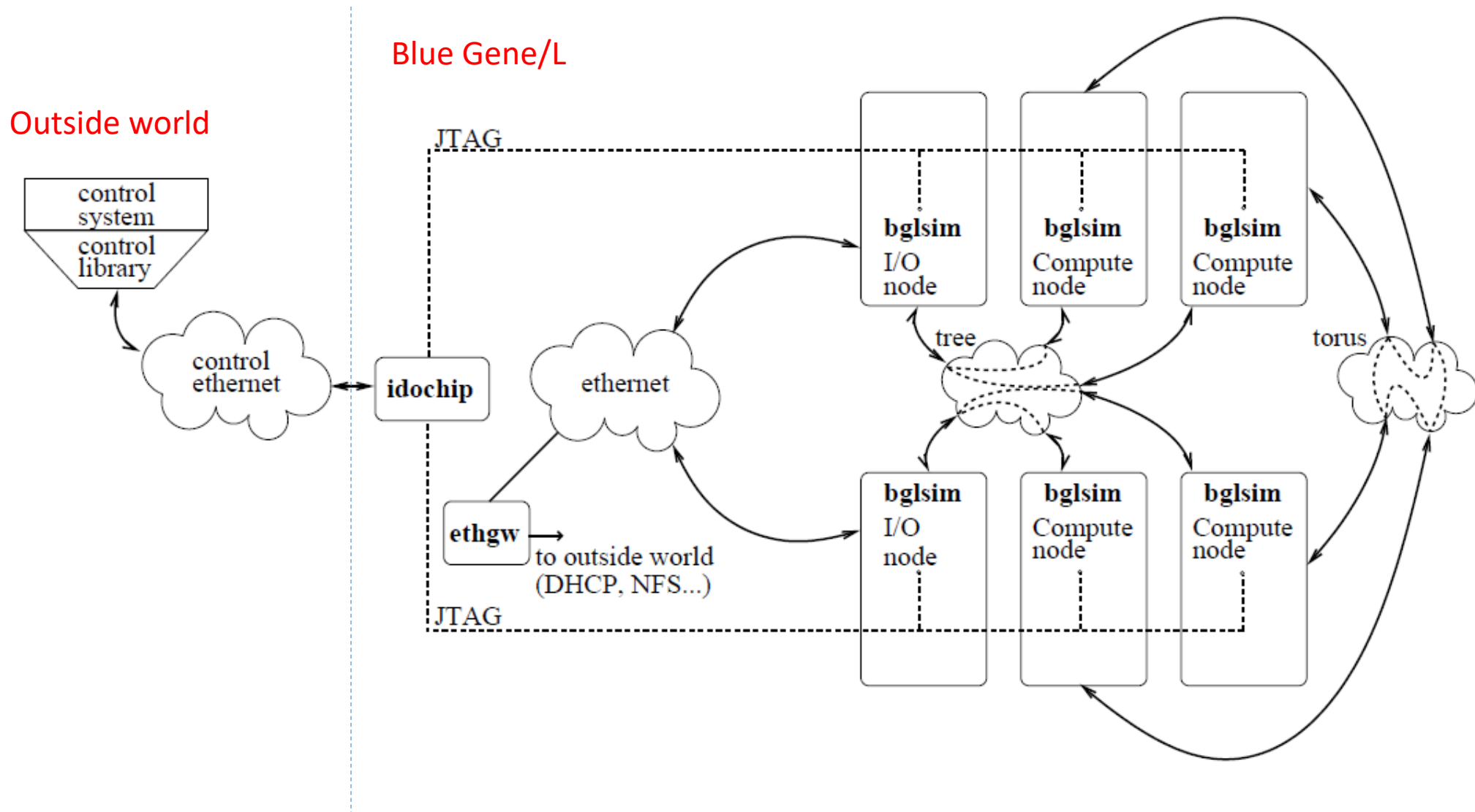
- Multi-node



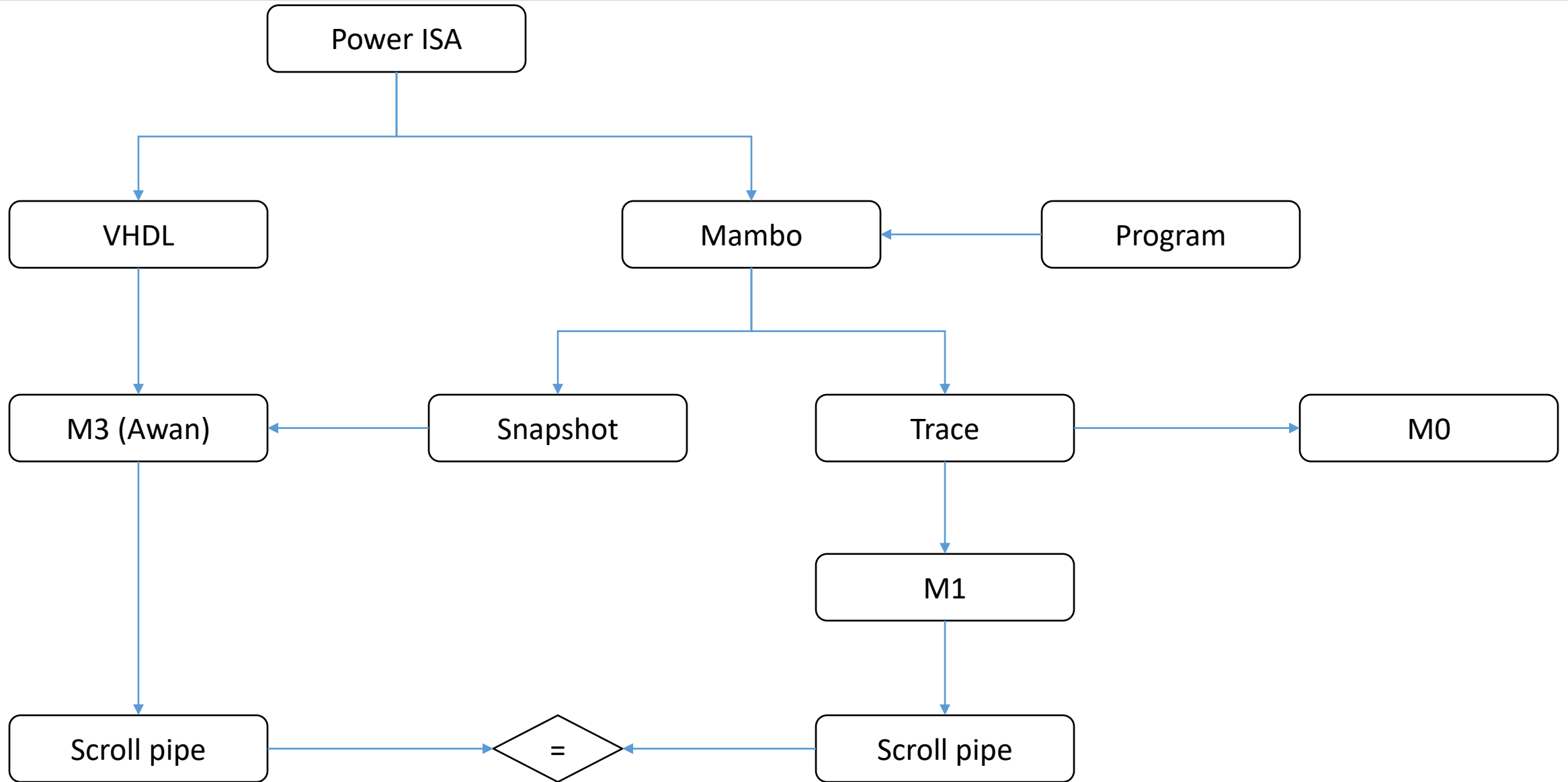
Mambo



Blue Gene/L full system simulation



Simulation process for current POWER processors



Overview of M1 and M3

M1

- M1 is the software model of the full processor chip
- It models most queues, branch predictors and caches but does not model any execution pipes (just latencies)
- M1 takes in the instruction trace (qtrace) generated by Mambo as input

M3

- M3 model is the real VHDL model that is loaded on Awan and run
- Available models for P10:
 - Core only – full SMT8 core + behavioral infinite L2
 - Chip – full SMT8 cores + L2 + L3 + nest (e.g., memory controller) + DIMM behavioral

Use the M1 model to project performance for a particular test and then use the M3 simulation models to verify that the VHDL matches the projection:

- M1 could be modeled inaccurately as well
- Could result in fixes in M1 and redefining targets

Awan overview

- VHDL cycle accurate simulator
- Specialized hardware
- Multiple concurrent simulations
- Shared by both POWER and Z teams

Frank Wallingford, *eDAW2021 Poster Session*



Comparing scroll pipes

M3

M1

The image displays a comparison of two scroll pipes, M3 and M1. The M3 view (left) shows a scroll pipe with various instructions and their completion times. The M1 view (right) shows a scroll pipe with various instructions and their completion times, including a list of instructions and their completion times.

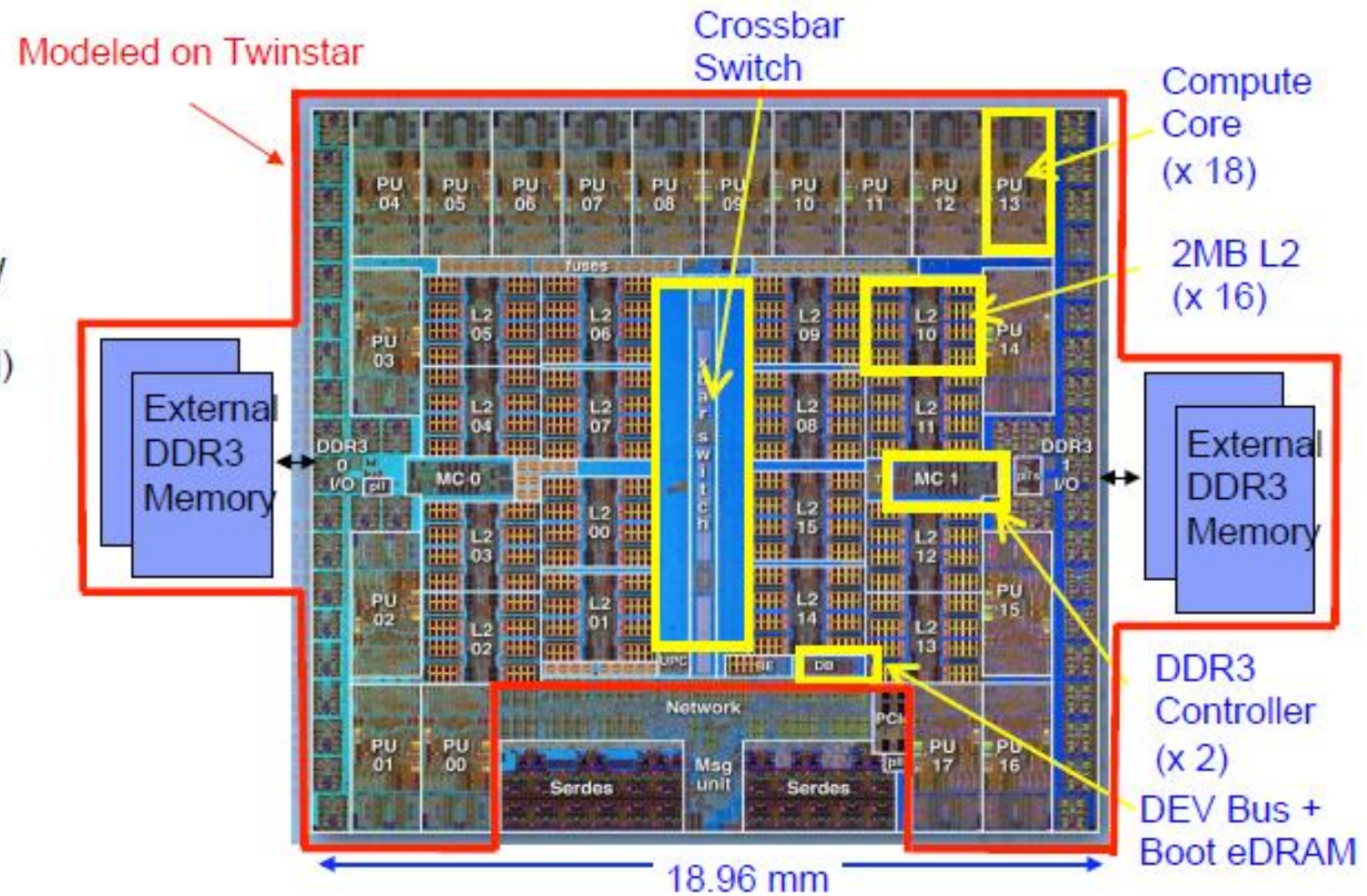
Instruction	Completion Time
D POSM	0:00007fff965f1ecc 0143 47951 stb r08,0(r10) 000
D POSM	0:00007fff965f1ecc 0143 1047951 stb r08,0(r10) 0
DP1SM	0:00007fff965f1ed0 0144 47960 lbz r10,0(r09) 000
DP1SM	0:00007fff965f1ed4 0145 47961 addi r08,r30,1
DdP1SM	0:00007fff965f1ed8 0146 47962 addi r09,r09,1
D P1SM	0:00007fff965f1edc 0147 47963 stb r10,0(r30) 000
D P1SM	0:00007fff965f1edc 0147 1047963 stb r10,0(r30) 0
ddP2SM	0:00007fff965f1ee0 0148 47964 lbz r06,0(r09) 000
P3SM	0:00007fff965f1ee4 014a 47966 addic. r07,r07,-1
P3SM	0:00007fff965f1ee8 014c 47968 addi r30,r08,1
DP2SM	0:00007fff965f1eec 014d 47969 or r10,r04,r04
DP2SM	0:00007fff965f1ef0 014e 47970 stb r06,0(r08) 000
DP2SM	0:00007fff965f1ef0 014e 1047970 stb r06,0(r08) 0
DP3SM	0:00007fff965f1ef4 014f 47971 beq 0x2c (bc 0xc,2,0x2c) [0x
D P2SM	0:00007fff965f1f20 0150 48140 cmpl cr7,r05,0
D P2SM	0:00007fff965f1f24 0151 48141 addi r31,r31,4
D P2SM	0:00007fff965f1f2c 0152 48142 beq cr7,-0x12c (bc 0xc,30,-0
dP2SM	0:00007fff965f1f28 0153 48143 or r04,r31,r31
P2SM	0:00007fff965f1f30 0154 48144 or r03,r30,r30
DP2SM	0:00007fff965f1f34 0155 48145 add r31,r31,r05
DP2SM	0:00007fff965f1f38 0156 48146 add r30,r30,r05
DdP2SM	0:00007fff965f1f3c 0157 48147 bl 0xffffef84 [0x7fff965f0e
D P2SM	0:00007fff965f0ec0 0158 48152 std r02,24(r01) 000
D P2SM	0:00007fff965f0ec0 0158 1048152 std r02,24(r01) 0
DDP2SM	0:00007fff965f0ec4 0159 48153 ld r12,-32496(r02) 000
D P1SM	0:00007fff965f0ec8 015a 48154 mtspr CTR,r12
D P1SM	0:00007fff965f0ecc 015b 48155 bctr (bcctr 0x15,0,0)
D POSMI0	0:00007fff96766d40 015c 48159 cmpli cr1,1,r05,31
D POSMI0	0:00007fff96766d44 015d 48160 neg r00,r03
DP1SM	0:00007fff96766d48 015e 48161 ble cr1,0x198 (bc 0x4,5,0x19
DP1SM	0:00007fff96766ee0 0160 48168 or r11,r03,r03
dP3SM	0:00007fff96766ee4 0161 48169 cmpli cr6,1,r05,8
P3SM	0:00007fff96766ee8 0162 48170 mtocrf 242,r05
P2SM	0:00007fff96766eec 0163 48171 ble cr6,0x104 (bc 0x4,25,0x1
DP0SM	0:00007fff96766ff0 0164 48172 bne cr6,-0x70 (bc 0x4,26,-0x
DP0SM	0:00007fff96766ff0 0166 48184 ble cr7,0x30 (bc 0x4,29,0x30
DP2SM	0:00007fff96766fb0 0168 48194 bne cr7,0x30 (bc 0x4,30,0x30
DDP0SM	0:00007fff96766fe0 016a 48306 bnsldr cr7,0 (bclr 0x4,31,0)
DDP0SM	0:00007fff96766fe4 016b 48307 lbz r06,0(r04) 000
DDP2SM	0:00007fff96766fe8 016c 48308 stb r06,0(r11) 000
DDP2SM	0:00007fff96766fe8 016c 1048308 stb r06,0(r11) 0
DDP2SM	0:00007fff96766fec 016d 48309 blr (bclr 0x15,0,0)

Research: modeling Blue Gene/Q compute node with FPGAs (2012)

- Proshanta Saha, Charles Haymes, Ralph Bellofalo, Bernard Brezzo, Mohit Kapur, Sameh Asaad

KEY CHIP FEATURES

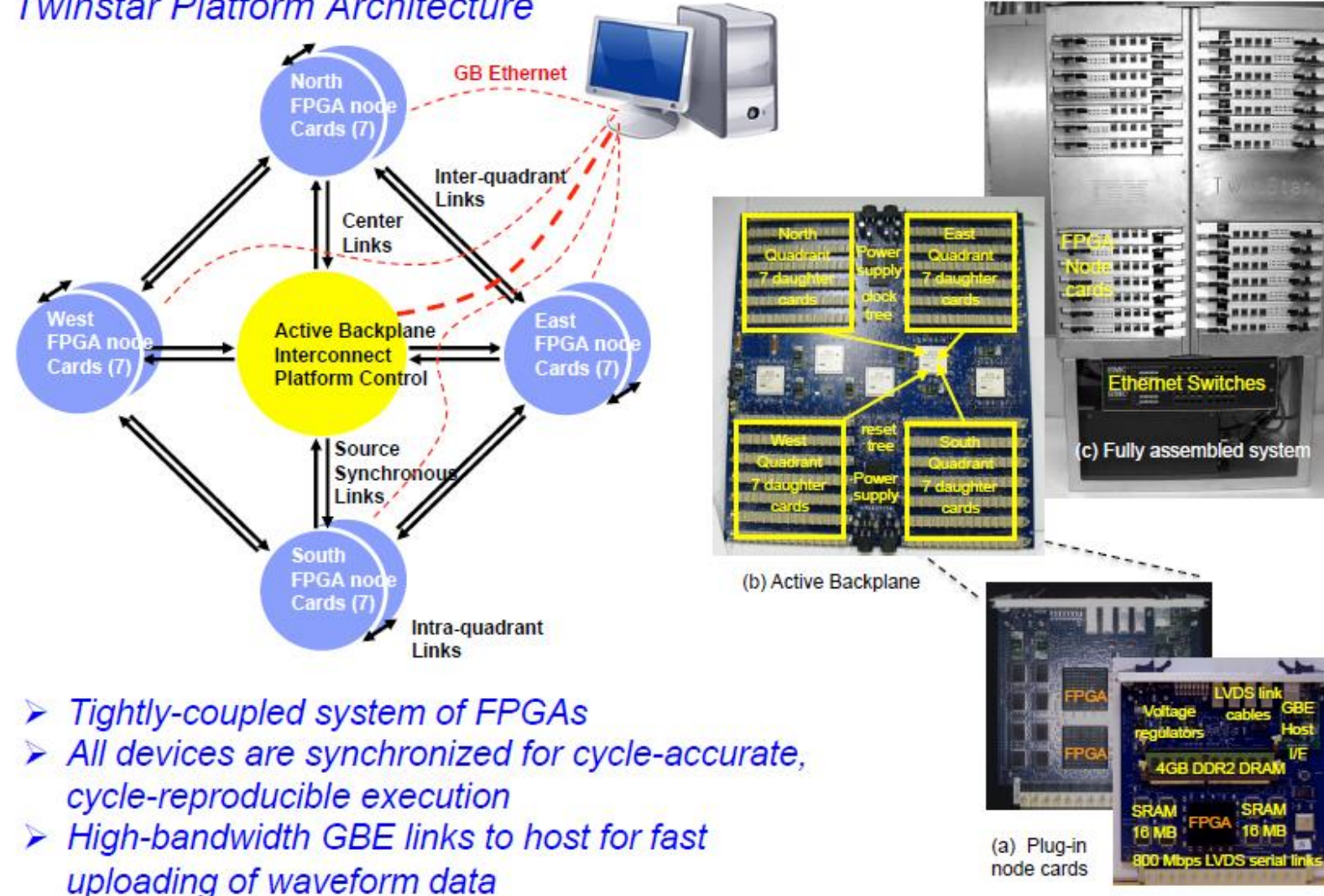
- 45-nm SOI technology
- 204.8 GFlops/sec @ 55W
- 17-way SMP
- 32 MB L2 cache (eDRAM)
- full crossbar switch
- Area: 360 mm²



RTL verification and debugging using FPGAs

- Proshanta Saha, Charles Haymes, Ralph Bellofalo, Bernard Brezzo, Mohit Kapur, Sameh Asaad

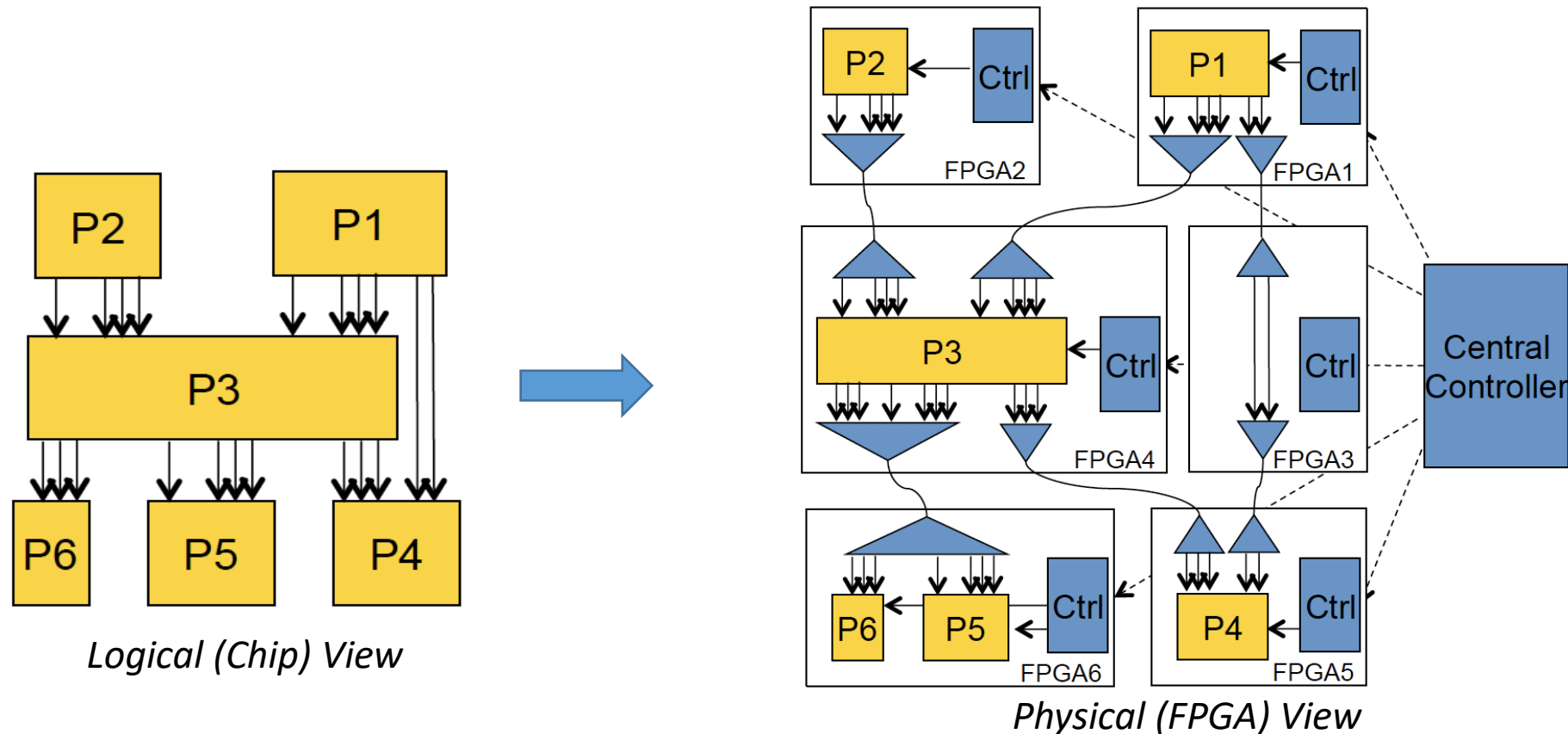
Twinstar Platform Architecture



- *Tightly-coupled system of FPGAs*
- *All devices are synchronized for cycle-accurate, cycle-reproducible execution*
- *High-bandwidth GBE links to host for fast uploading of waveform data*

Partitioning onto multiple FPGA devices

- Proshanta Saha, Charles Haymes, Ralph Bellofalo, Bernard Brezzo, Mohit Kapur, Sameh Asaad



- One-to-one and many-to-one mapping of chip partitions onto individual FPGA devices
- Serializer & de-serializer links inserted between partitions to fit physical connectivity
- Special treatment of clock and reset networks
- Chip partitions are stoppable, infrastructure logic is free-running

Conclusions

- We use a variety of simulation tools in the development of our large systems (Blue Gene, POWER, Z)
- Mambo has been in use for 20+ years – it was essential for Blue Gene and it is essential now, playing a variety of roles
- Performance modeling relies on a variety of simulators, with increasing complexity and fidelity: M0, M1, M3
- VHDL-level simulation (M3) requires hardware acceleration
- Most *exploration* today happens with M0 or M1