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)

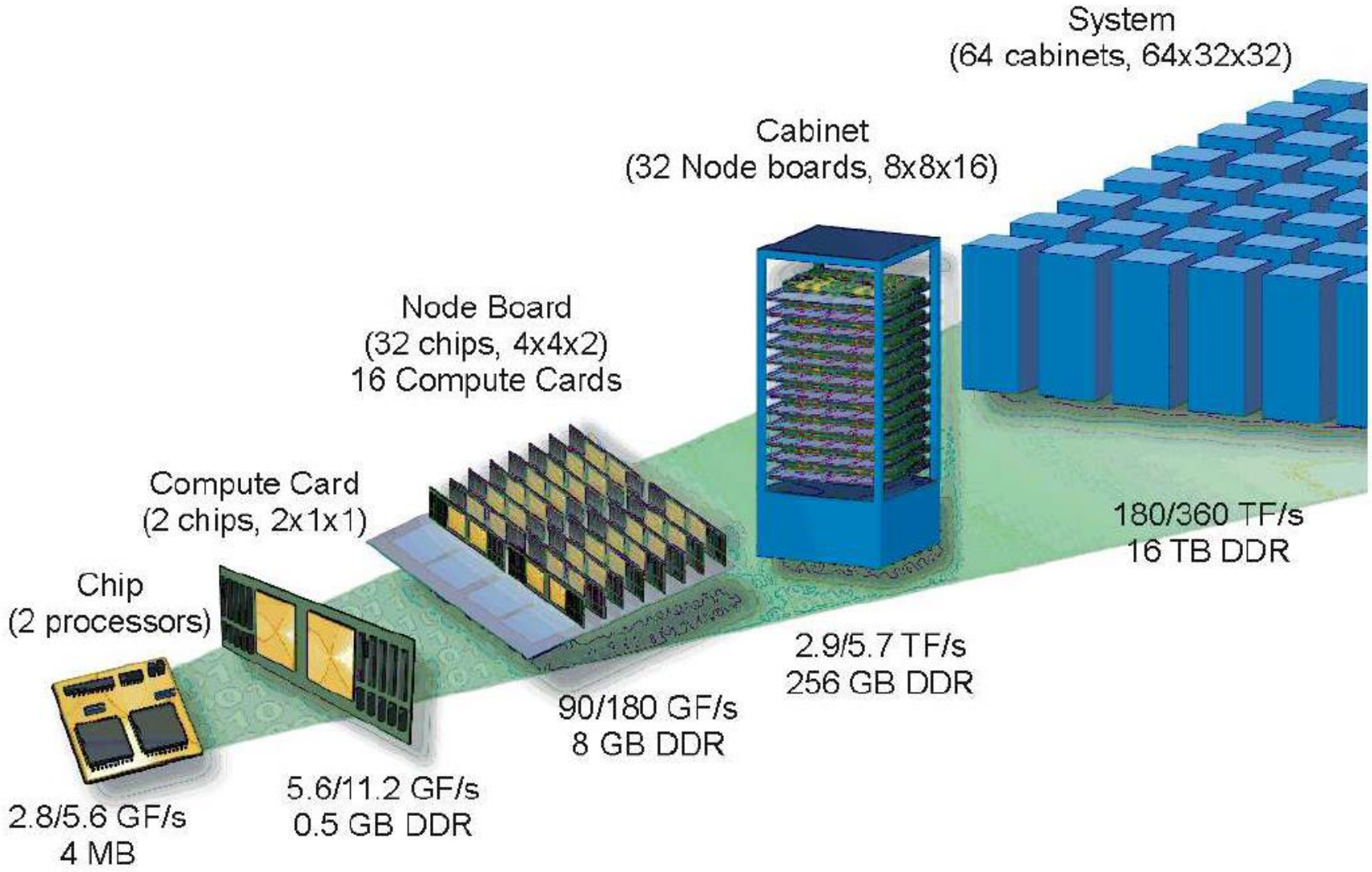
Motherhood and apple pie …

- 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 …

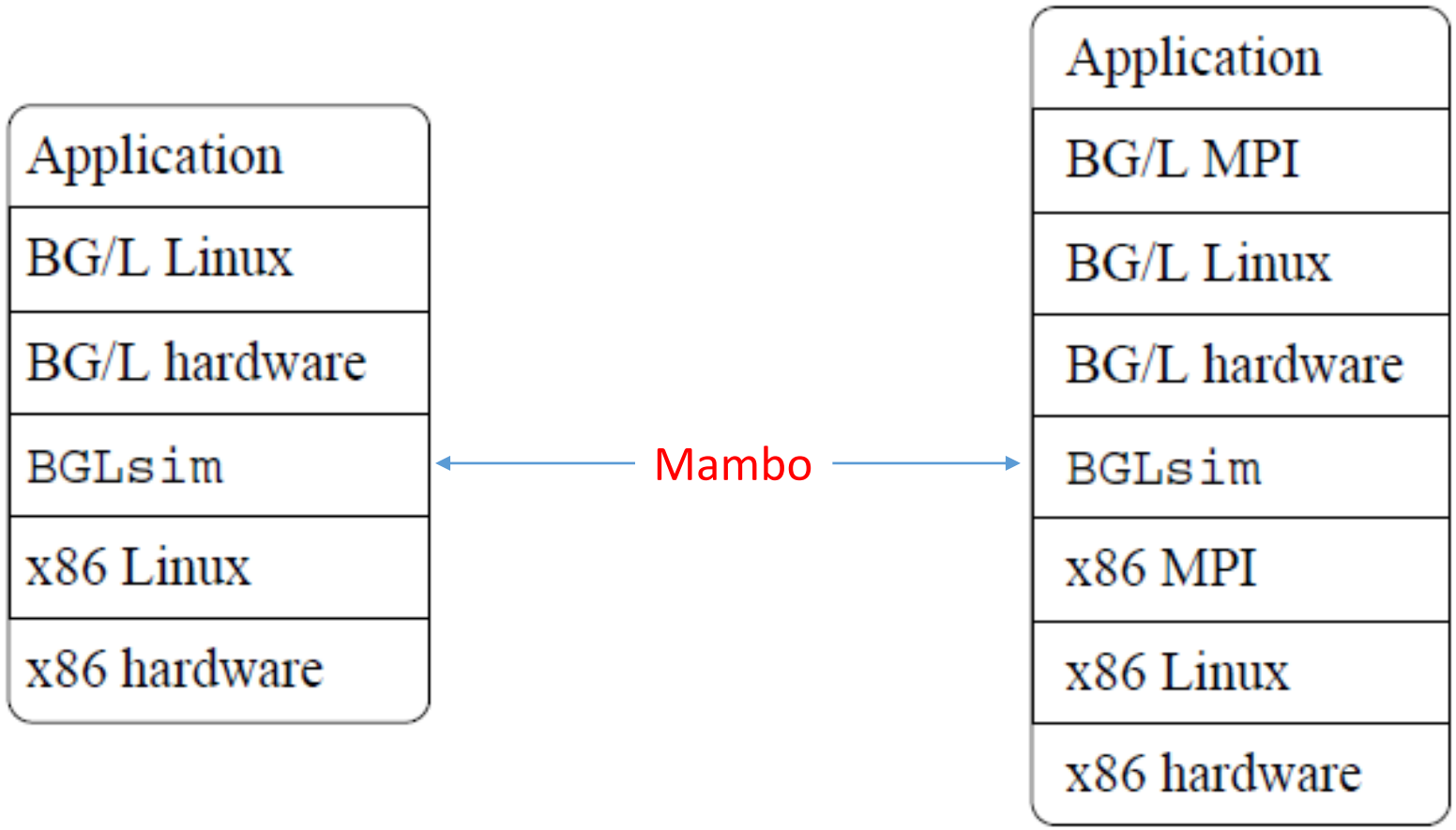
- Mambo (also called **`systemsim`**) is an ISA-level functional simulator
- A C program that runs on x86 and Power ISA systems
- It models a multi-core system with shared memory
- It serves multiple official purposes:
  - A reference ISA implementation, used for testing the design
  - A software development 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)

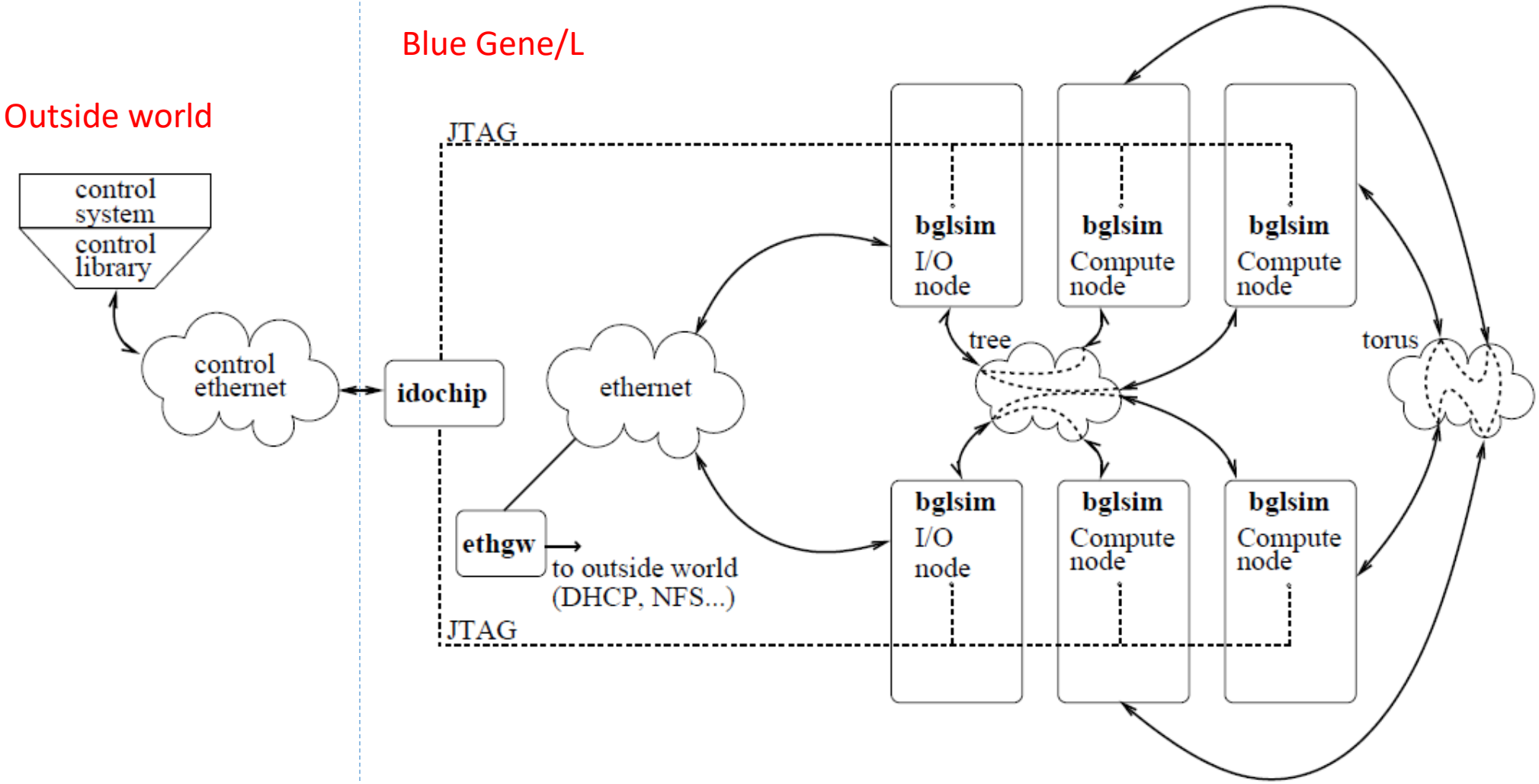


The Blue Gene/L system simulation stack

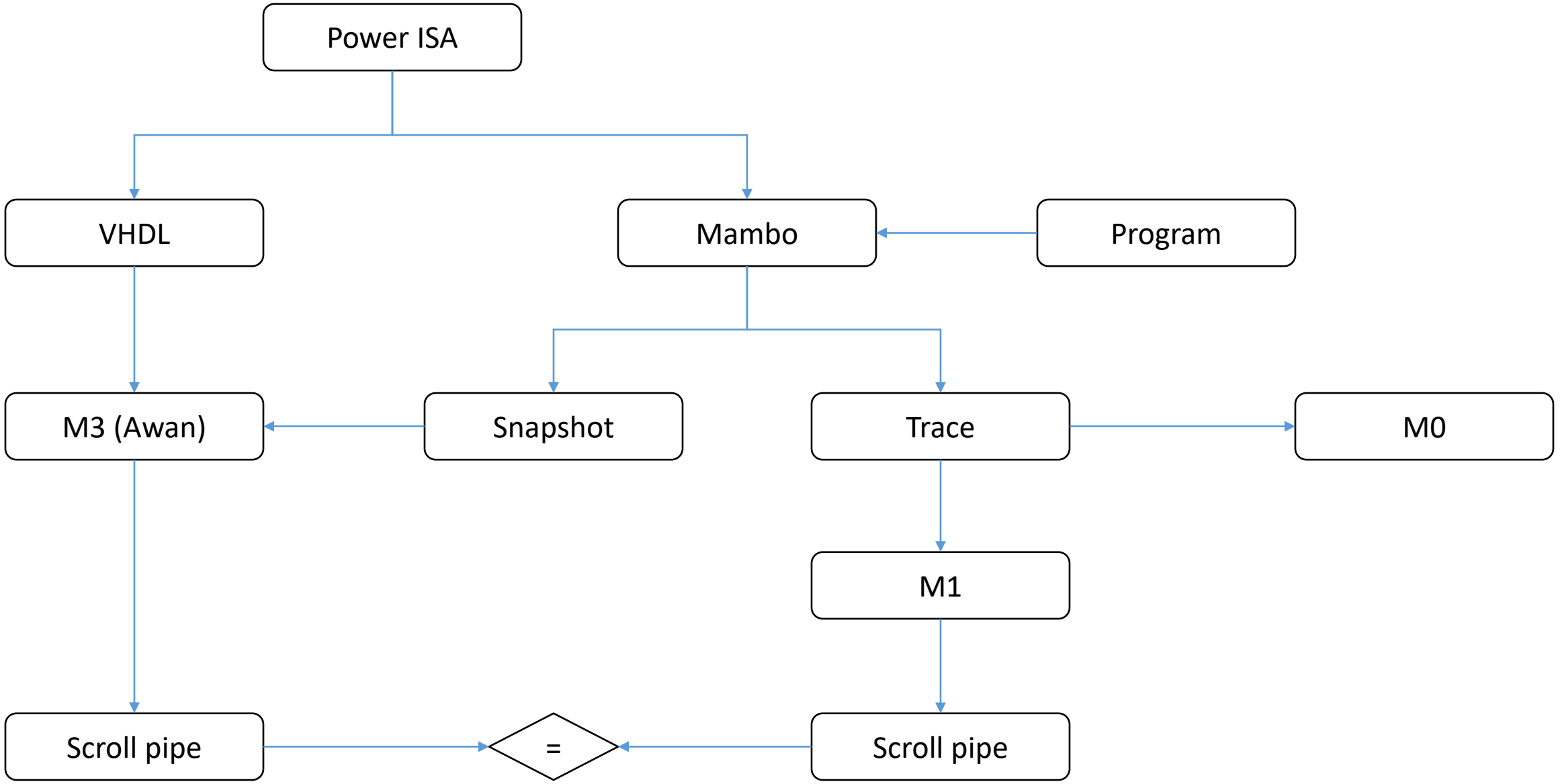
- Single-node
- Multi-node



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

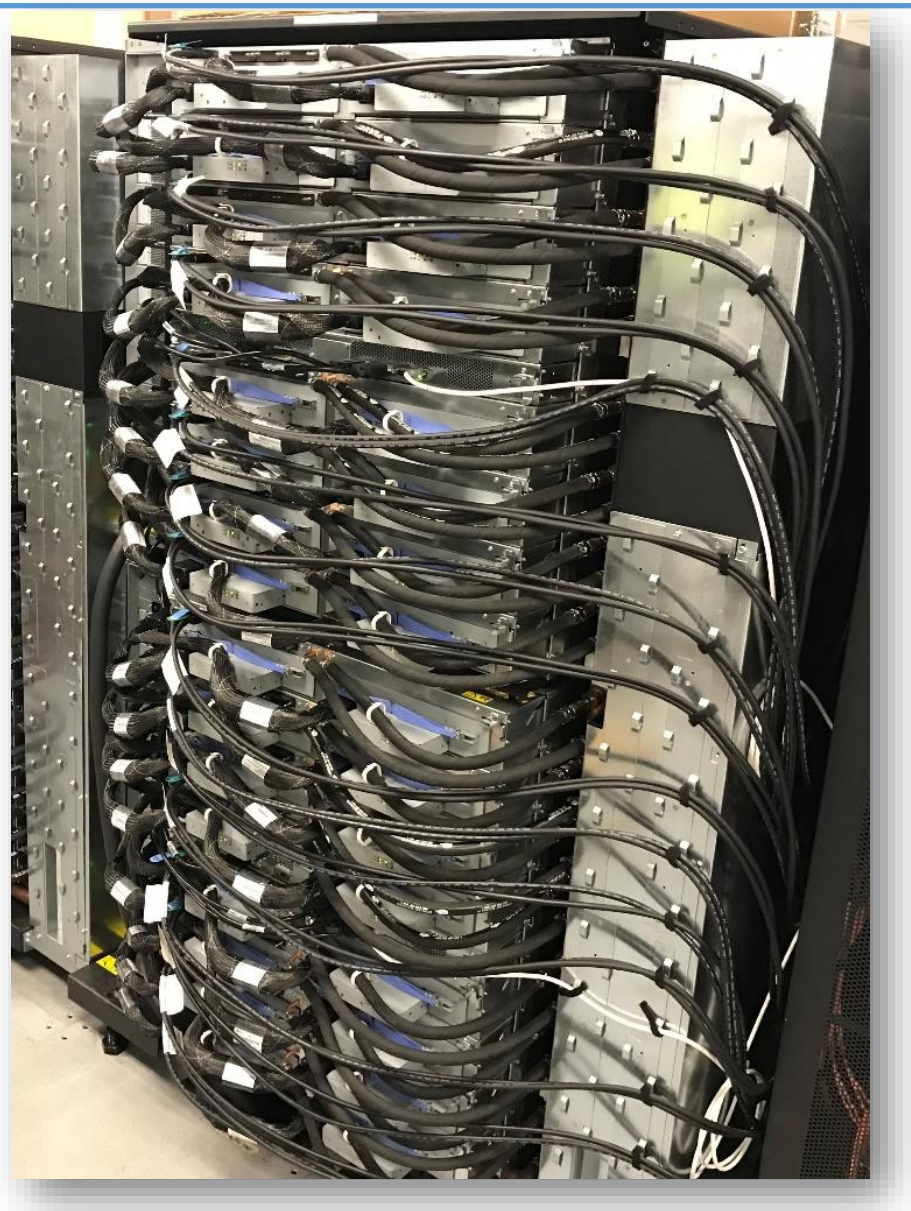
- M3 model is the real VHDL model that is loaded on Awan and run
- Available models for P10:
  - Core only – full SMT8 core + behavioral infinite L2
  - Chip – full SMT8 cores + L2 + L3 + nest (e.g., memory controller) + DIMM behavioral

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

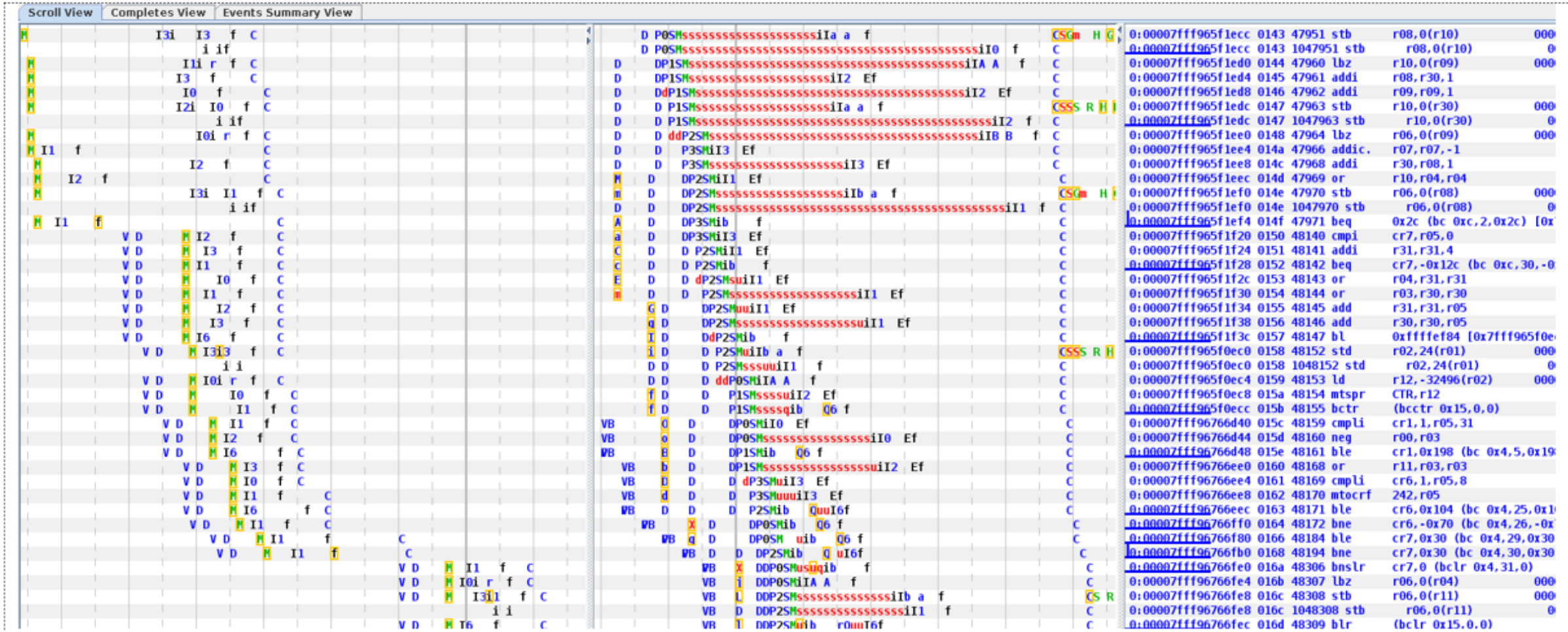
- VHDL cycle accurate simulator
- Specialized hardware
- Multiple concurrent simulations
- Shared by both POWER and Z teams



Frank Wallingford, *eDAW2021 Poster Session*

Comparing scroll pipes

M3 M1

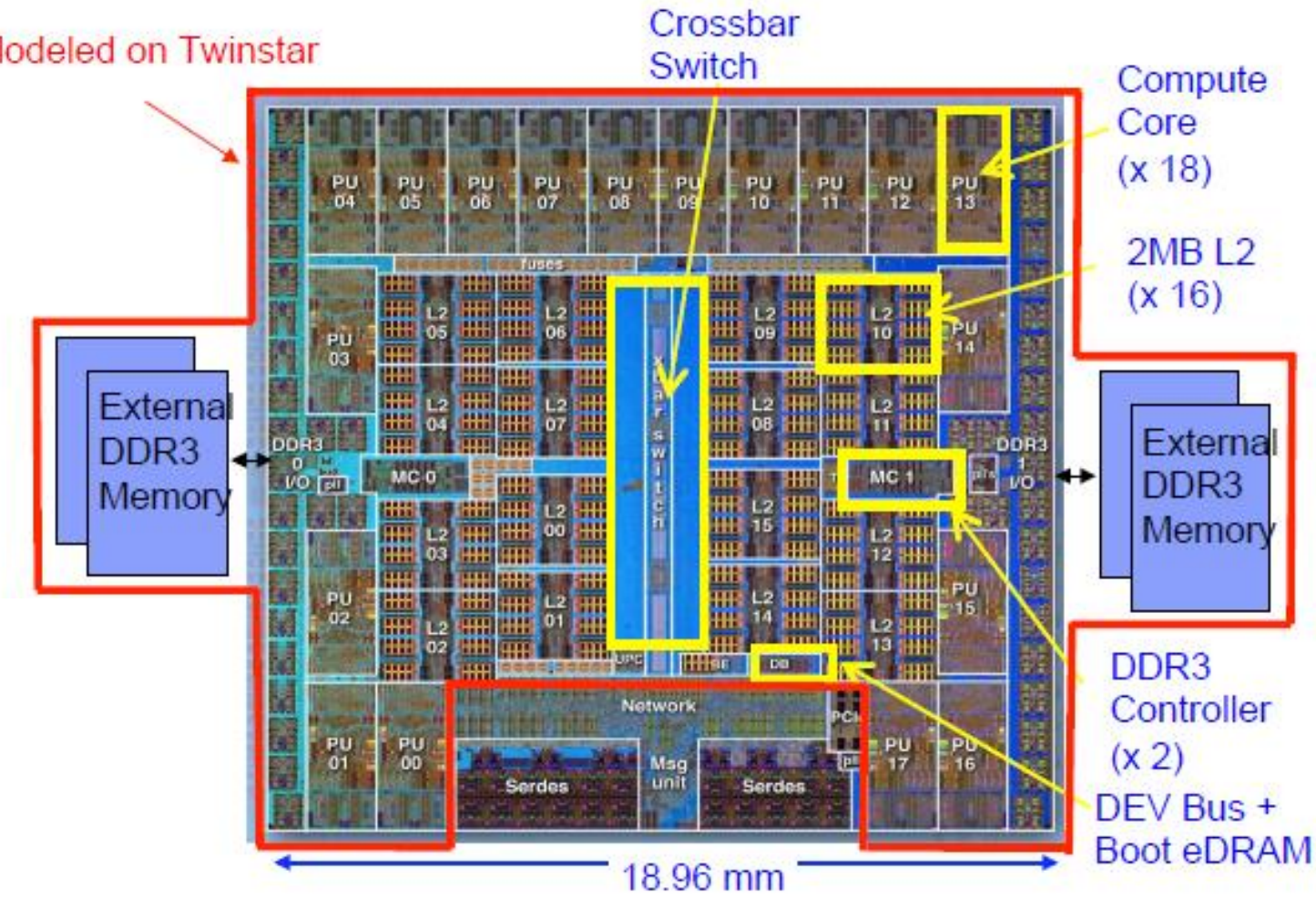


Research: modeling Blue Gene/Q compute node with FPGAs (2012)

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

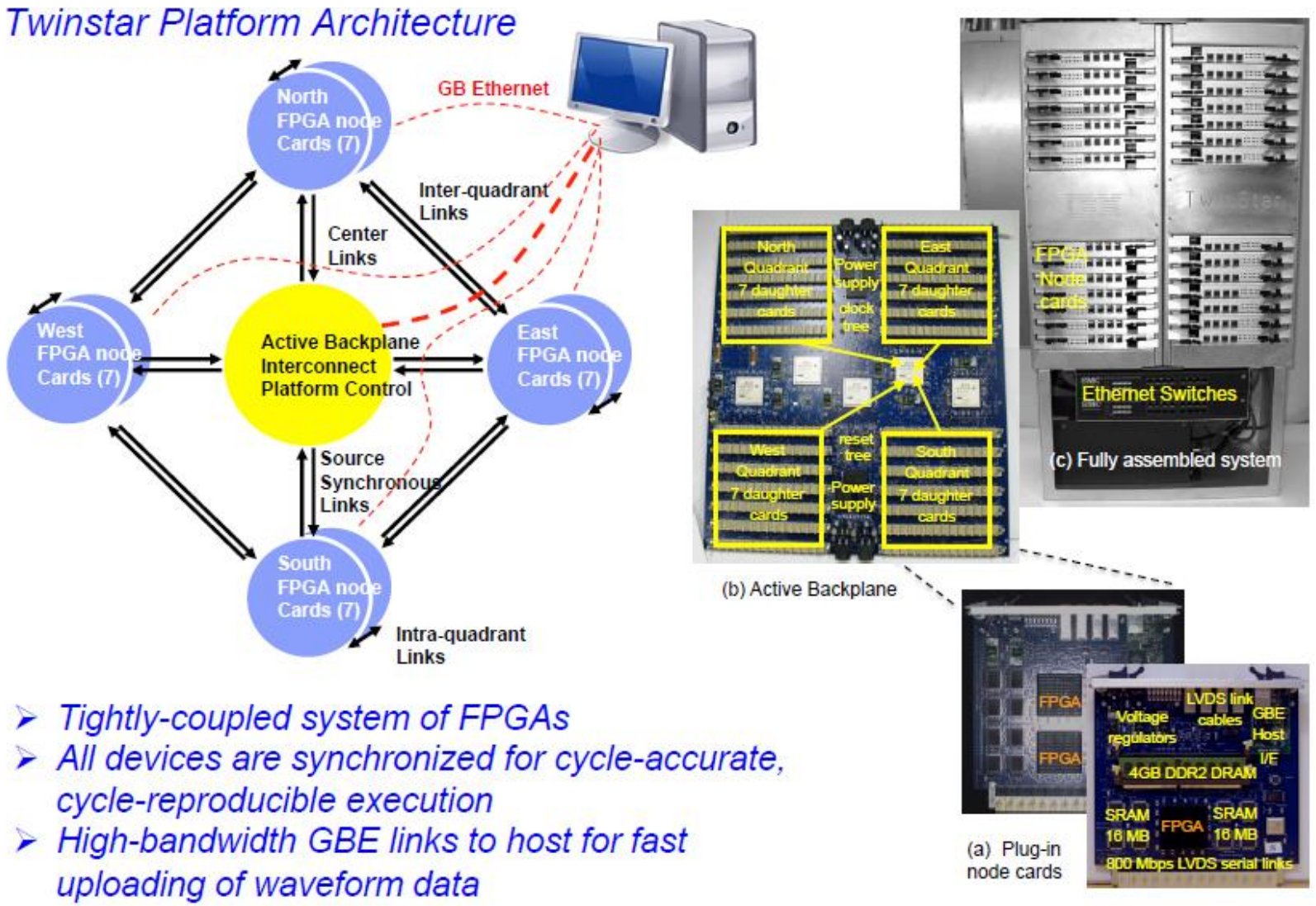
KEY CHIP FEATURES

- 45-nm SOI technology
- 204.8 GFlops/sec @ 55W
- 17-way SMP
- 32 MB L2 cache (eDRAM)
- full crossbar switch
- Area: 360 mm^2



RTL verification and debugging using FPGAs

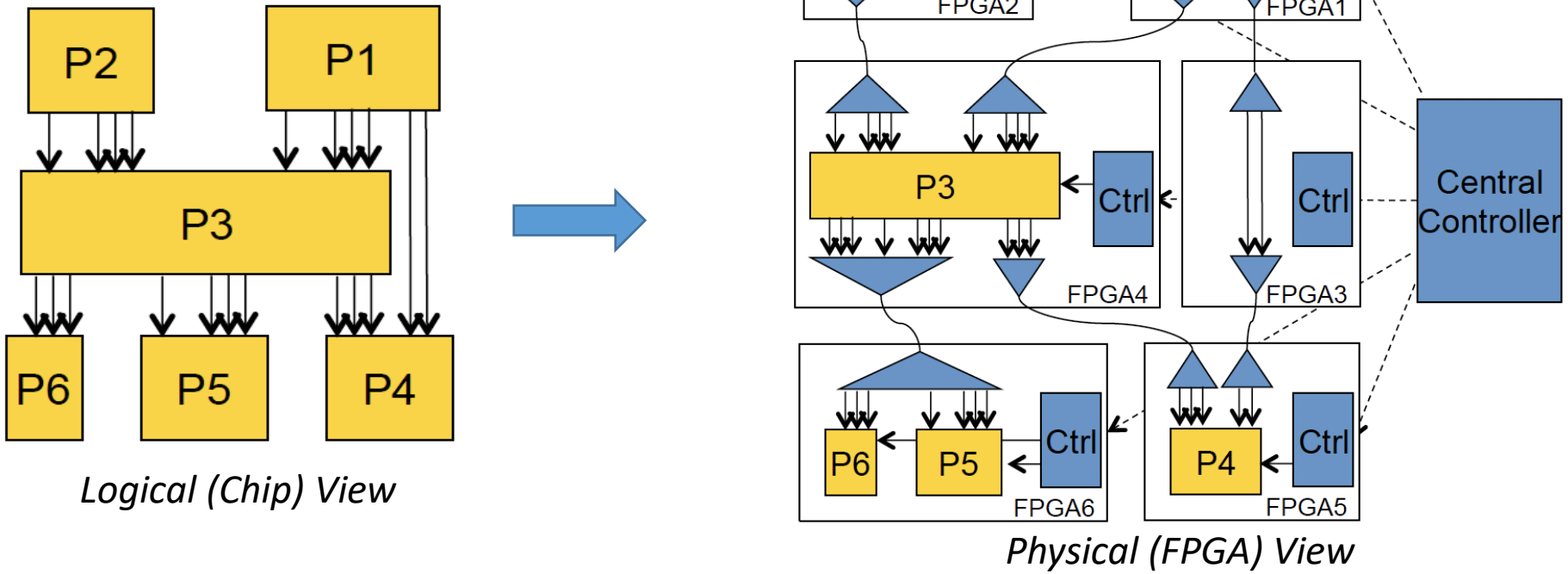
- Proshanta Saha, Charles Haymes, Ralph Bellofalo, Bernard Brezzo, Mohit Kapur, Sameh Asaad



- *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