A 5.2 GHz Microprocessor Chip for the IBM zEnterprise™ System

J. Warnock[1], Y. Chan[2], W. Huott[2], S. Carey[2], M. Fee[2], H. Wen[3], M.J. Saccamango[2], F. Malgioglio[2], P. Meaney[2], D. Plass[2], Y.-H. Chan[2], M. Mayo[2], G. Mayer[4], L. Sigal[5], D. Rude[2], R. Averill[2], M. Wood[2], T. Strach[4], H. Smith[2], B. Curran[2], E. Schwarz[2], L. Eisen[3], D. Malone[2], S. Weitzel[3], P.-K. Mak[2], T. McPherson[2], C. Webb[2]

IBM Systems and Technology Group:

1 - Yorktown Heights, NY

2 - Poughkeepsie, NY

3 – Austin, TX

4 - Boeblingen, Germany

5 IBM Research, Yorktown Heights, NY

Outline

- Introduction
- Technology and chip overview
  - RAS
  - Clock grids
- Core circuit design
- L3 design
- Power and noise considerations
- Chip frequency tuning
- Conclusions

Introduction: zEnterprise 196 (z196)

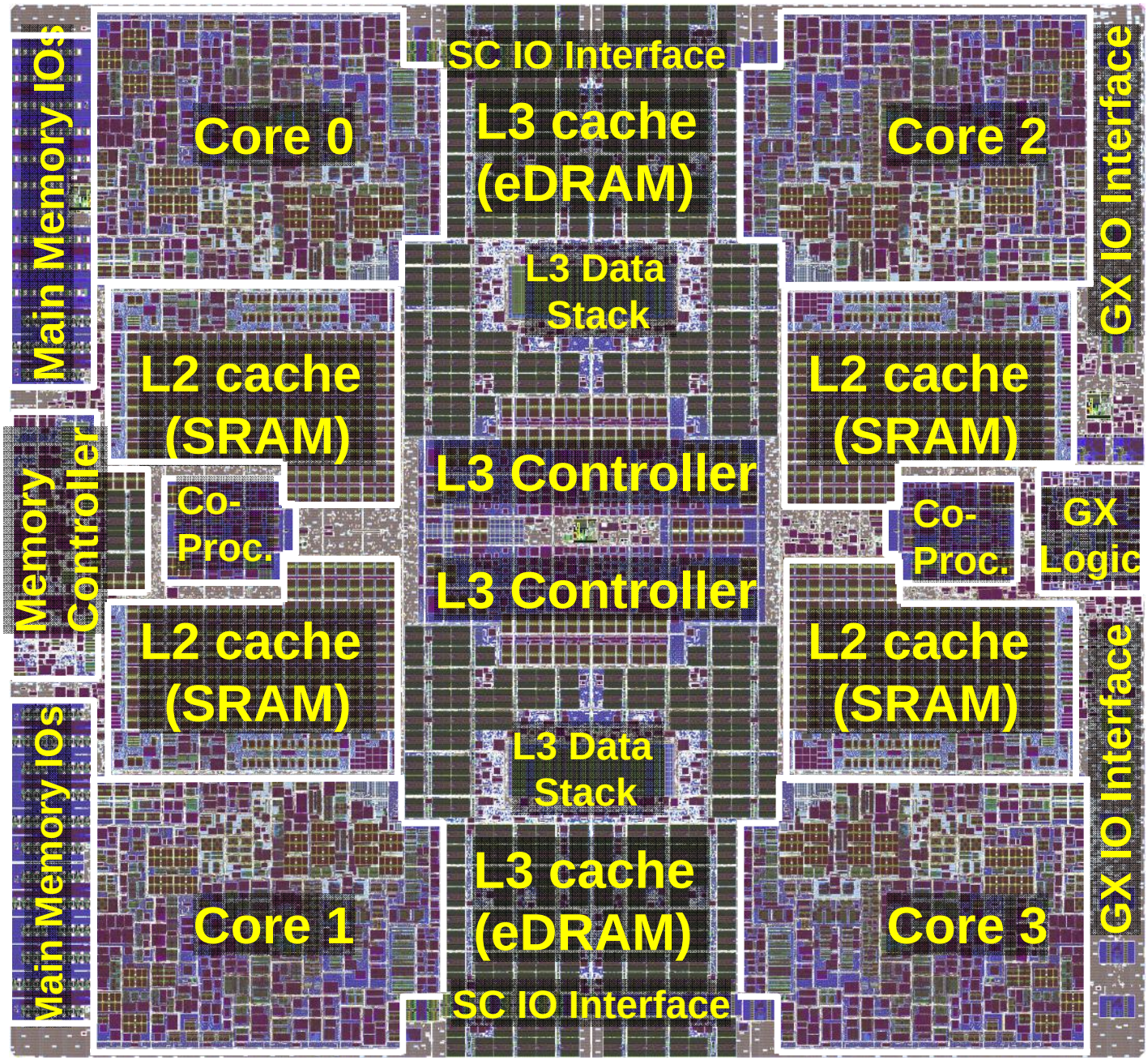
- Single-thread performance is critical for system z applications
- Starting point for z196: high-frequency z10 design
  - Z10 cores run at 4.4 GHz, 65nm technology
  - Move to 45nm technology for z196
  - Maintain low-FO4 pipeline for maximum frequency boost
  - Add out-of-order execution: improved IPC
  - Design improvements to reduce power dissipation
  - Improvements to 4-level cache hierarchy
- Result: Achieved 5.2 GHz operating frequency
  - Same power envelope as previous system
  - Up to 40% improvement seen on legacy workloads

Outline

- Introduction
- Technology and chip overview
  - RAS
  - Clock grids
- Core circuit design
- L3 design
- Power and noise considerations
- Chip frequency tuning
- Conclusions

Chip Technology & Design Overview

- High-performance 45nm SOI technology
- Embedded DRAM
  - Deep trench decoupling capacitors available
- Four logic threshold voltage options
  - Low V_T option for ultra-critical paths
- 2 Extra high-performance wiring planes added
- 4 cores, 1.5MB L2/core, 24MB shared L3
- Two co-processors, DDR3 RAIM controller, I/O bus controller (GX)
- 1.4B Transistors, 512mm^2 chip area


Main Memory IOs
SC IO Interface
Core 0
L3 cache (eDRAM)
Core 2
GX IO Interface
L3 Data Stack
L2 cache (SRAM)
L2 cache (SRAM)
Memory Controller
Co-Proc.
L3 Controller
Co-Proc.
GX Logic
L3 Controller
L2 cache (SRAM)
L2 cache (SRAM)
Main Memory IOs
L3 Data Stack
GX IO Interface
Core 1
L3 cache (eDRAM)
Core 3
SC IO Interface

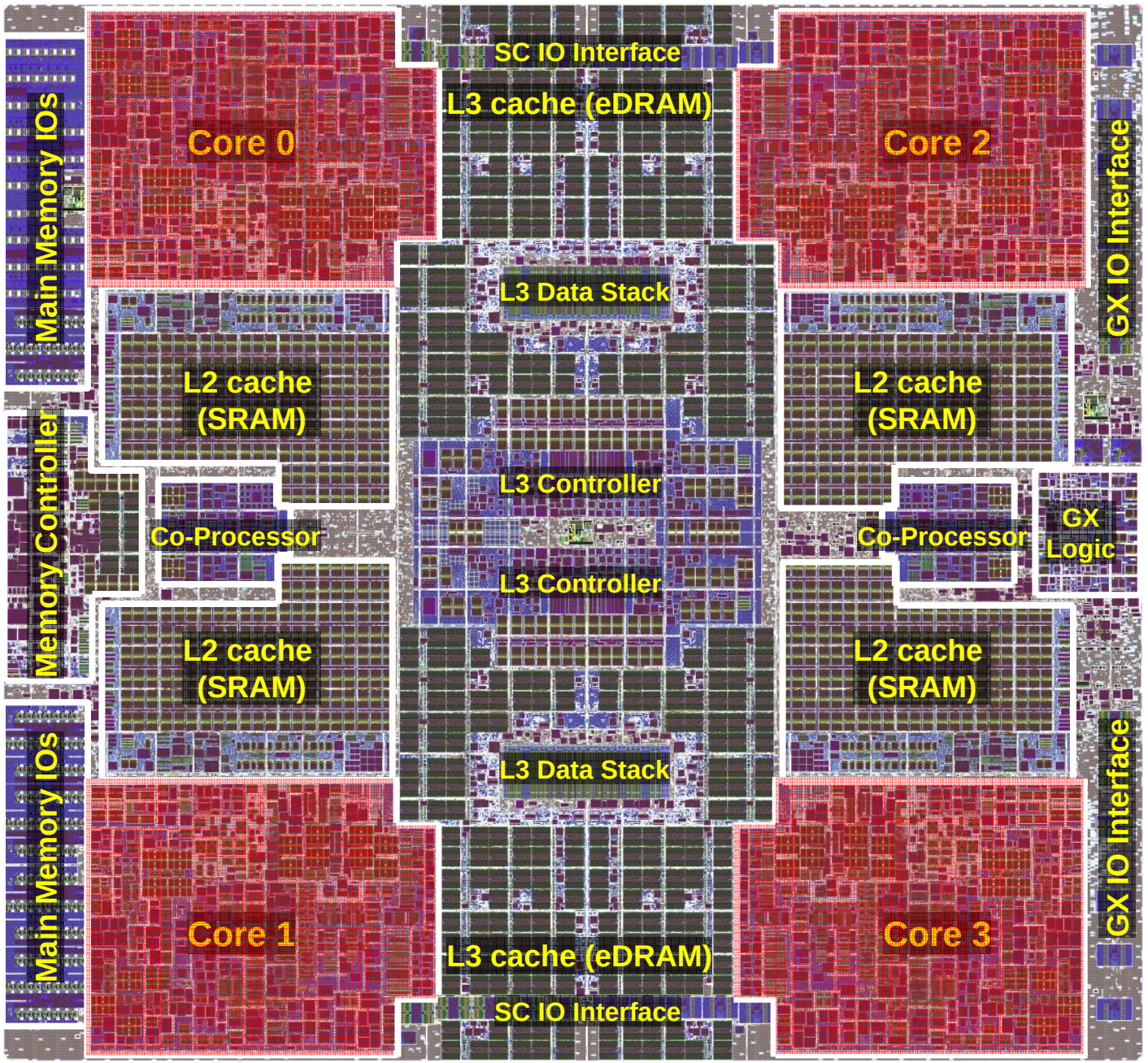
Chip RAS Features

- Base technology: SOI provides SER advantage
- Component-level hardening against SER
  - Stacked devices in most clocked storage elements
- Extensive circuit-level techniques used
  - Parity, residues, local duplication of function
  - Checking overhead estimated at 20-25% for digital logic
- On-chip caches: ECC, parity

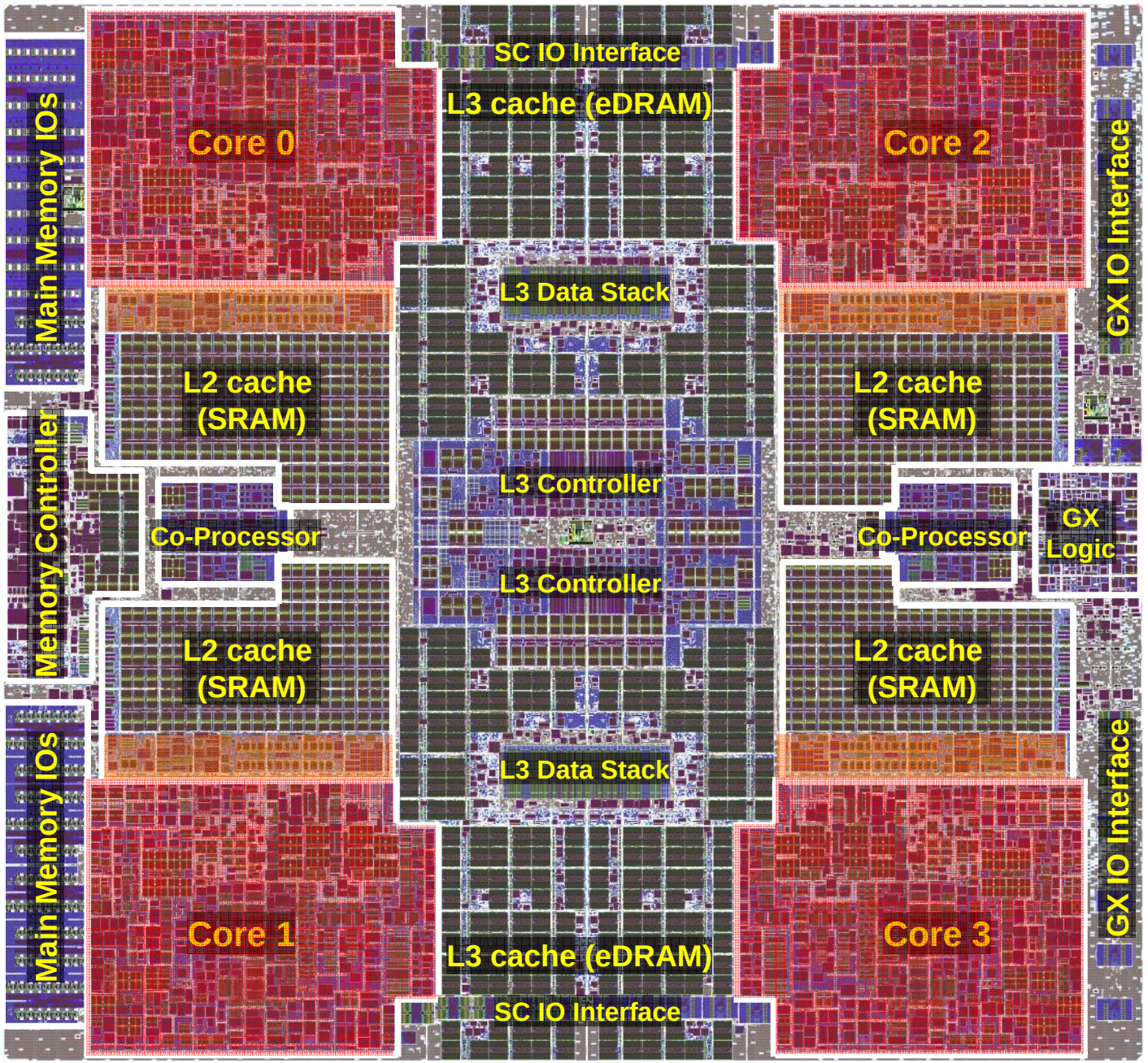
  Memory: RAIM ECC, Bus CRC, Tiered Recovery
- Recovery Unit (RU)
  - Maintains checkpointed states
  - RU restarts processor from checkpointed state when error detected

Chip Clock Grids

- High frequency grids over each core
  - “1:1” clock period, 5.2 GHz
  - Grid extends over L2 control region (circuits run at 2:1 gear ratio)
- Half-speed grid over L2 & nest: “2:1 grid”: 2.6 GHz
  - Most of nest is geared down by a factor of 2 = 4:1
- Synchronous interfaces from core to L2
  - 1:1 -> 2:1, on separate grids
- Synchronous interfaces from L2 to L3
  - 2:1 -> 4:1, but on the same grid
- Separate asynchronous grids over I/O region
  - Some overlap of synchronous and asynchronous grids

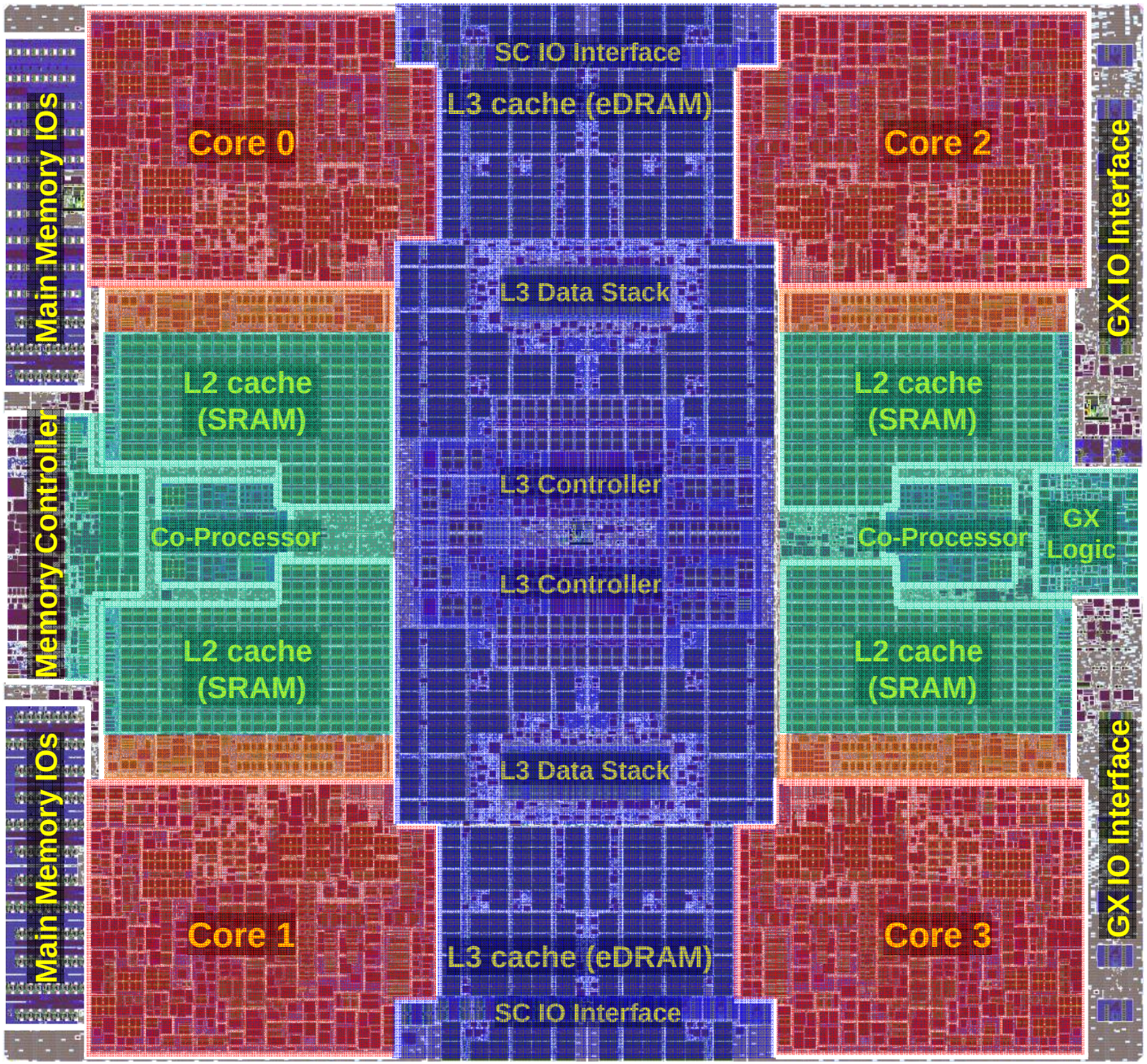

SC IO Interface
L3 cache (eDRAM)
Core 0
Core 2
Main Memory IOs
GX IO Interface
L3 Data Stack
L2 cache (SRAM)
L2 cache (SRAM)
Memory Controller
L3 Controller
Co-Processor
Co-Processor
GX Logic
L3 Controller
L2 cache (SRAM)
L2 cache (SRAM)
Main Memory IOs
L3 Data Stack
GX IO Interface
Core 1
Core 3
L3 cache (eDRAM)
SC IO Interface

1:1 grid



1:1 grid

1:1 grid
(Circuits at 2:1)

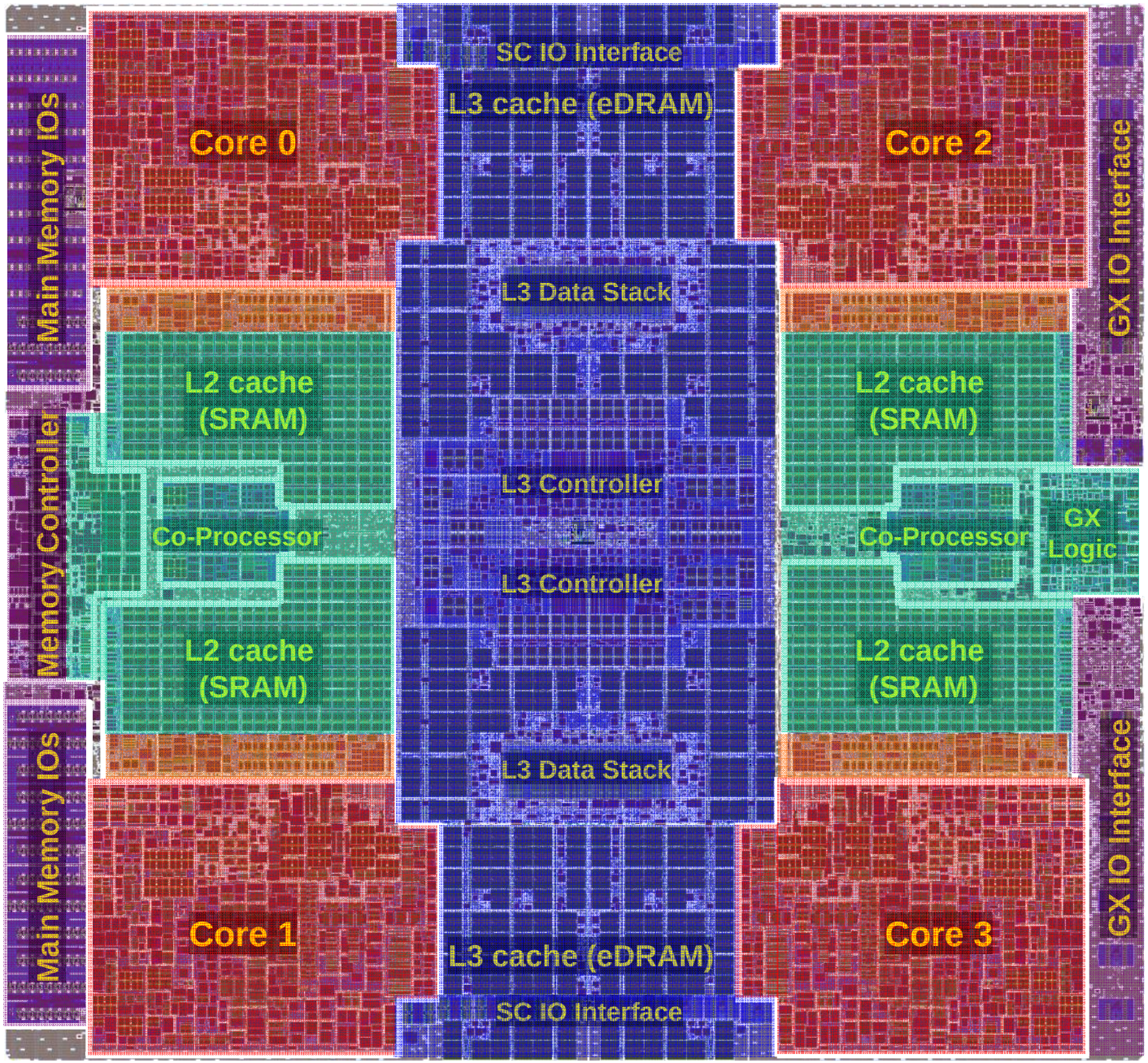


1:1 grid

1:1 grid
(Circuits at 2:1)

2:1 grid

2:1 grid
(Circuits at 4:1)



1:1 grid

1:1 grid
(Circuits at 2:1)

2:1 grid

2:1 grid
(Circuits at 4:1)

Asynch grids

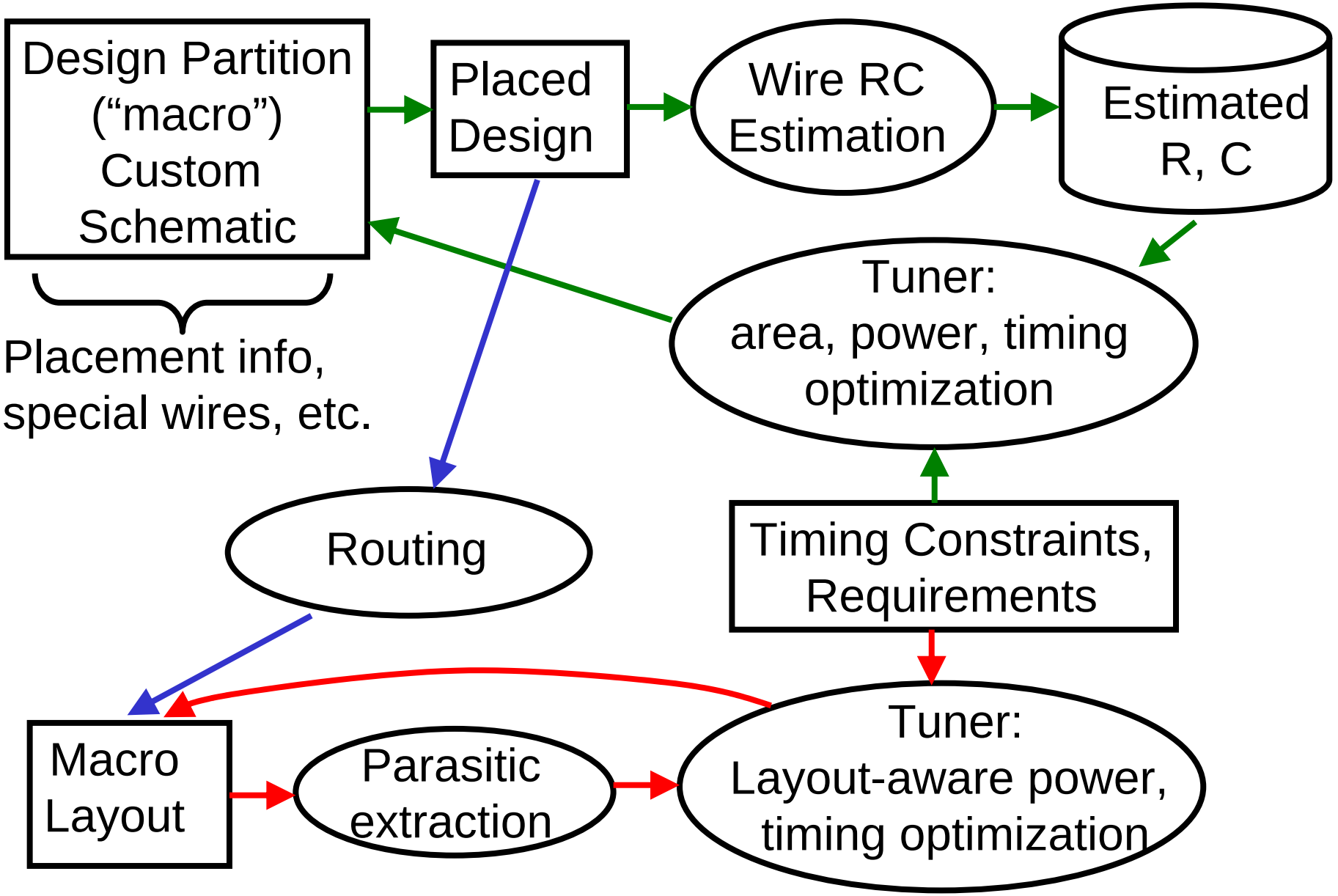
Outline

- Introduction
- Technology and chip overview
  - RAS
  - Clock grids
- Core circuit design
- L3 design
- Power and noise considerations
- Chip frequency tuning
- Conclusions

Processor Core Circuit Design

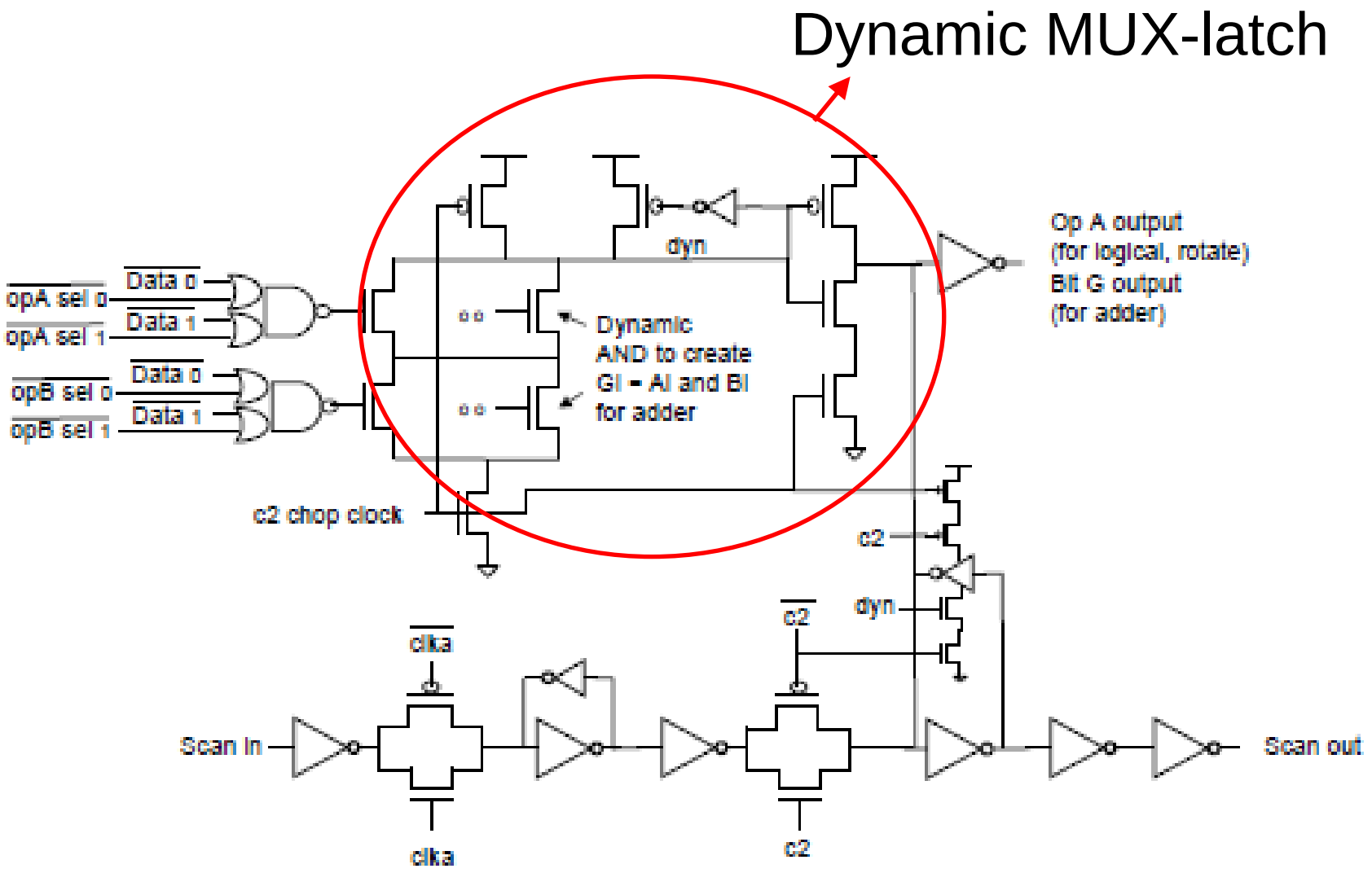
- Custom dataflow implementation
  - Static CMOS design with parameterized gates
  - Automated device width, VT tuning (pre- and post-layout)
  - Aggressive use of pulsed local clocks for power savings
  - Widespread fine-grained local clock gating
- Custom high-speed memory elements
  - 64KB I cache, 128 KB D cache
  - Dynamic circuits for critical access paths
- Synthesized control logic
  - Structured placement of clocked storage elements
  - Custom-like techniques for robust pulsed-clock routing

Custom Design Methodology



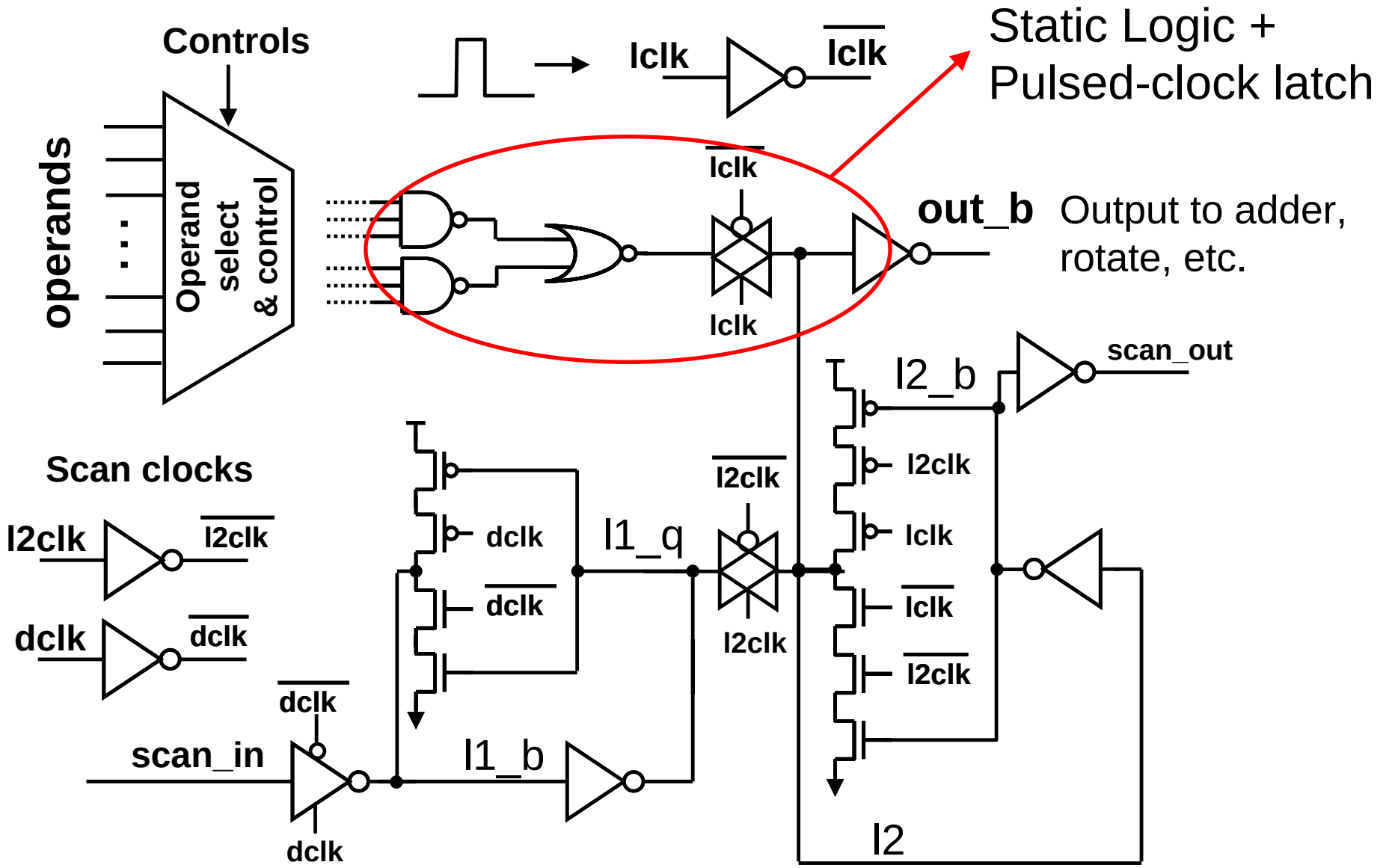
Single-cycle FXU Execution Loop

(Previous Generation Design)

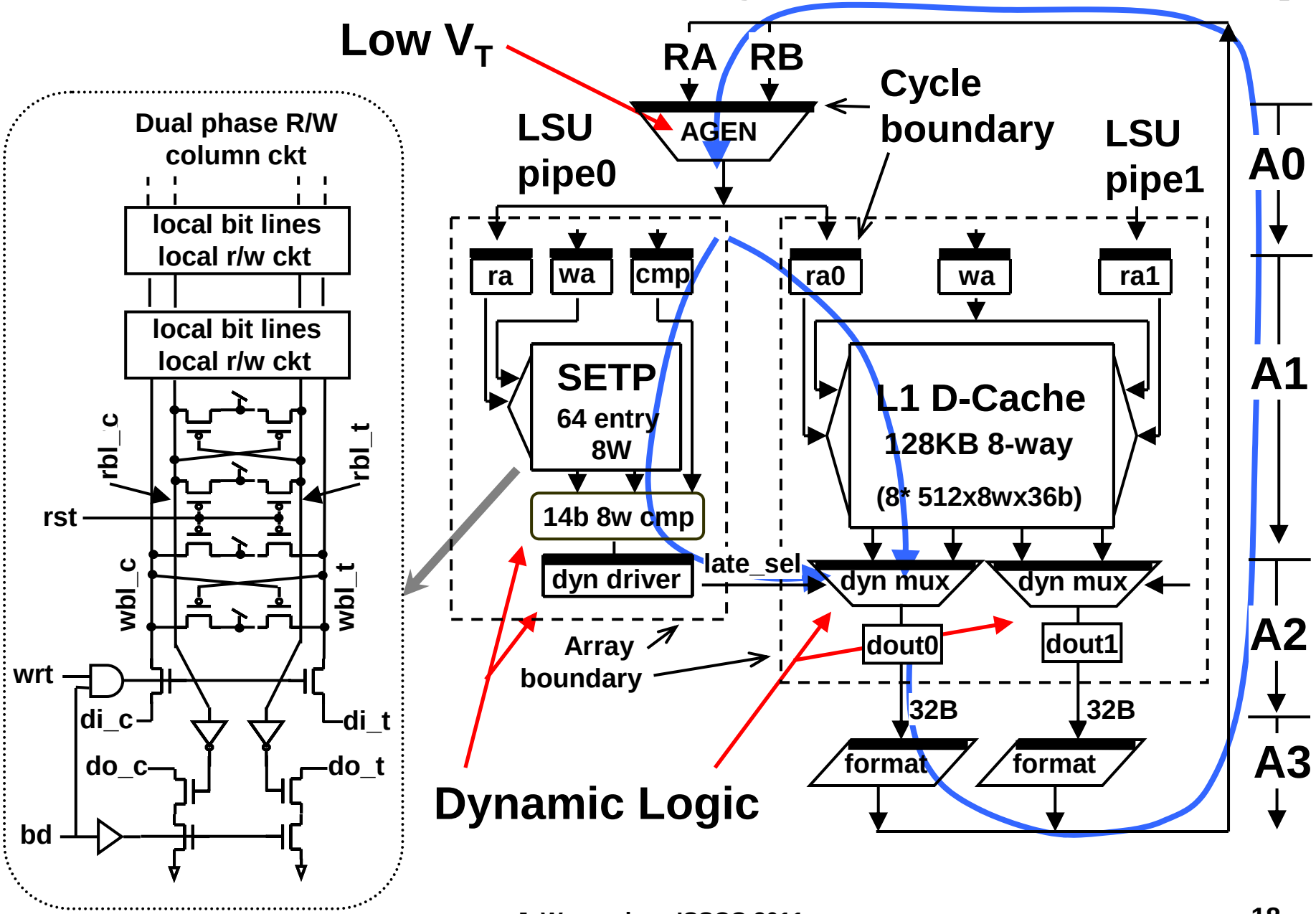


Single-cycle FXU Execution Loop

(Current Design)



L1 D-cache Critical 4-cycle Access Loop



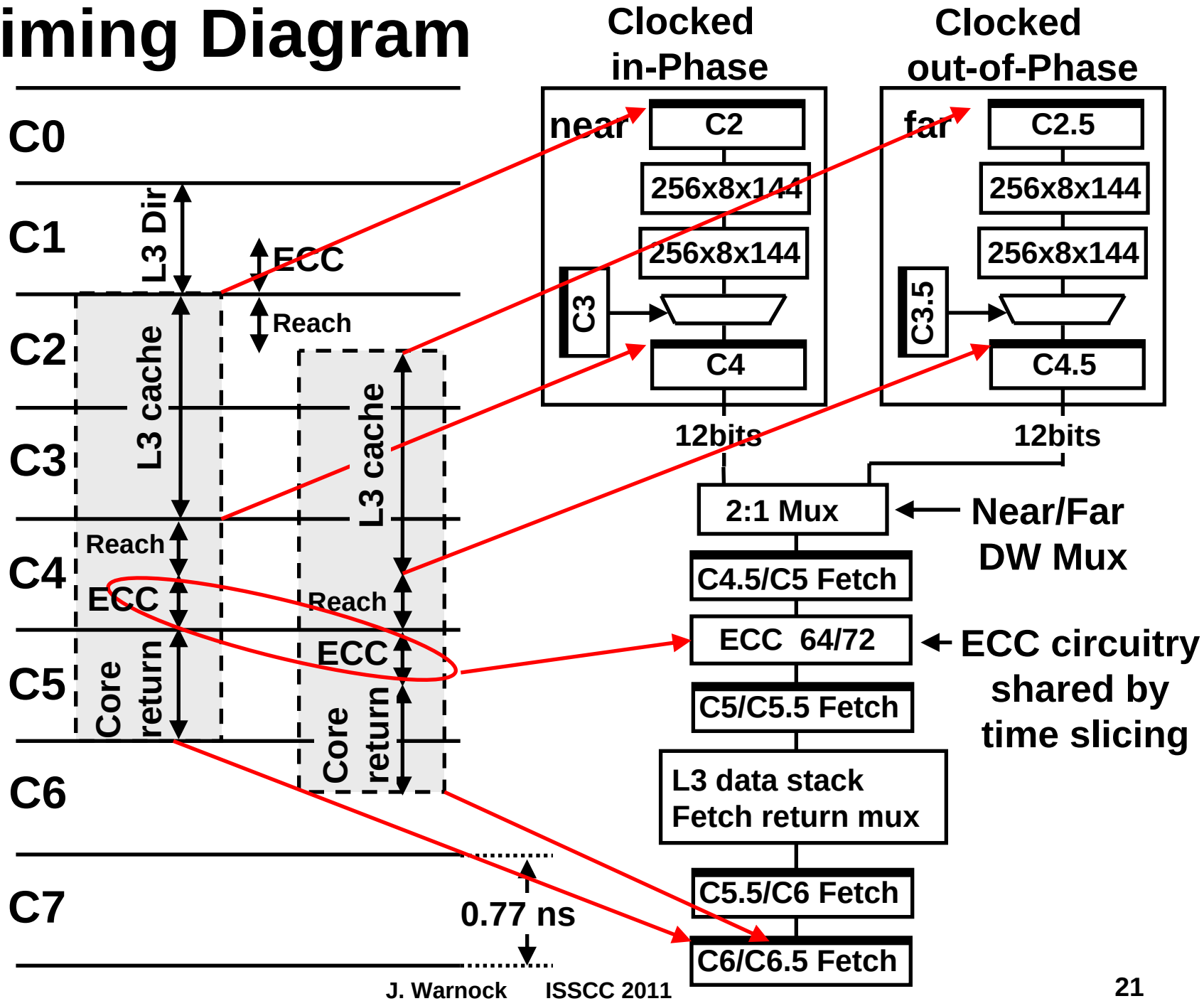
Outline

- Introduction
- Technology and chip overview
  - RAS
  - Clock grids
- Core circuit design
- L3 design
- Power and noise considerations
- Chip frequency tuning
- Conclusions

L3 Design

- 24MB on-chip shared L3 cache
  - Uses high-performance DRAM macros
  - 1.54ns access time for individual DRAM macro
  - 196 MB 4th level cache (L4) on separate SC chip
- Equal L3 access time from any of the 4 cores
  - 45 core processor cycles for L3 hit
  - Near/far data clocked on alternate 2:1 cycles
  - Allows sharing of circuitry for near/far data
- Lower level caches are store-through
  - Drives significant store activity in the L3
  - => highly interleaved design

L3 Timing Diagram



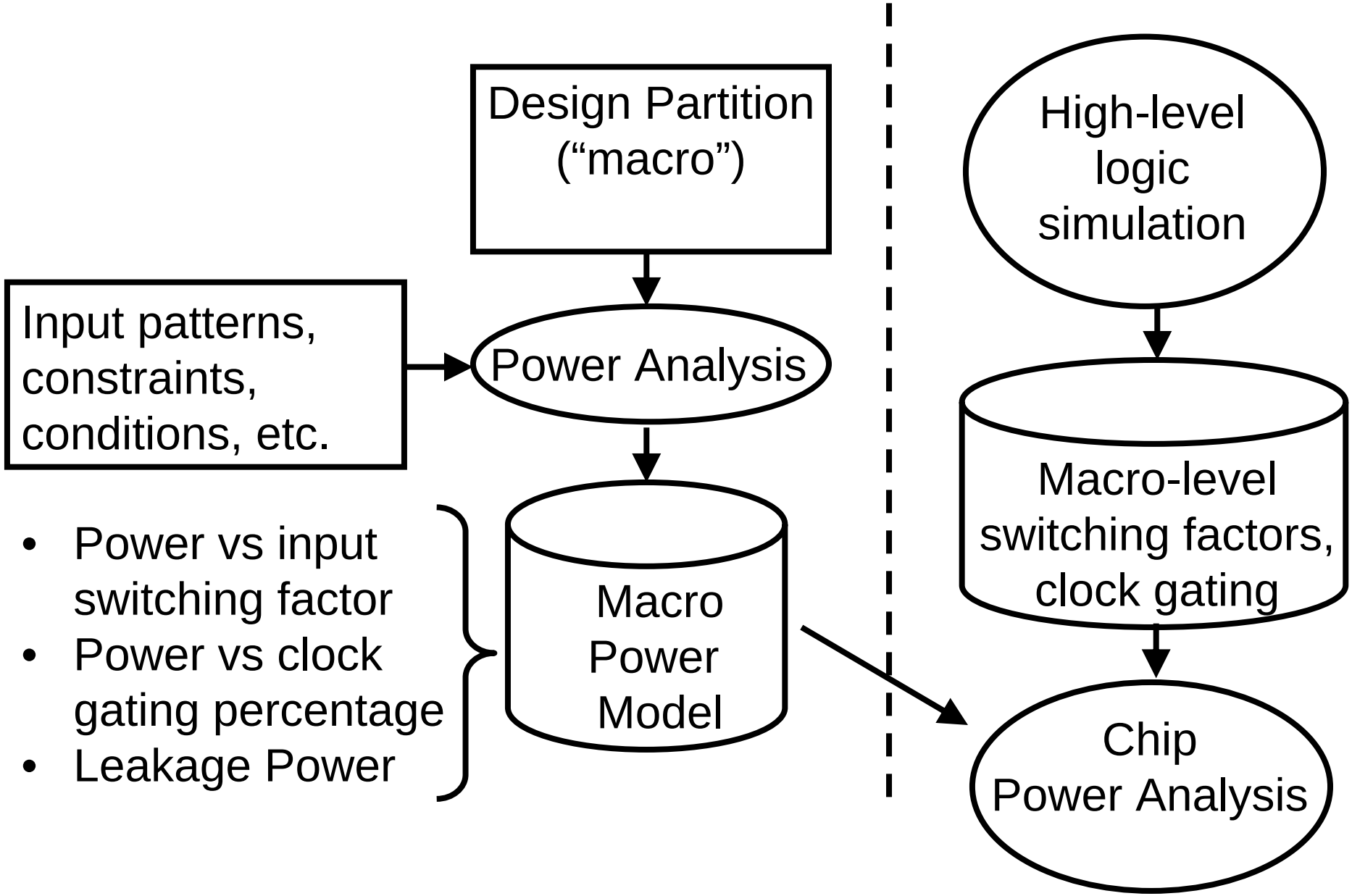
Outline

- Introduction
- Technology and chip overview
  - RAS
  - Clock grids
- Core circuit design
- L3 design
- Power and noise considerations
- Chip frequency tuning
- Conclusions

Chip Power Considerations

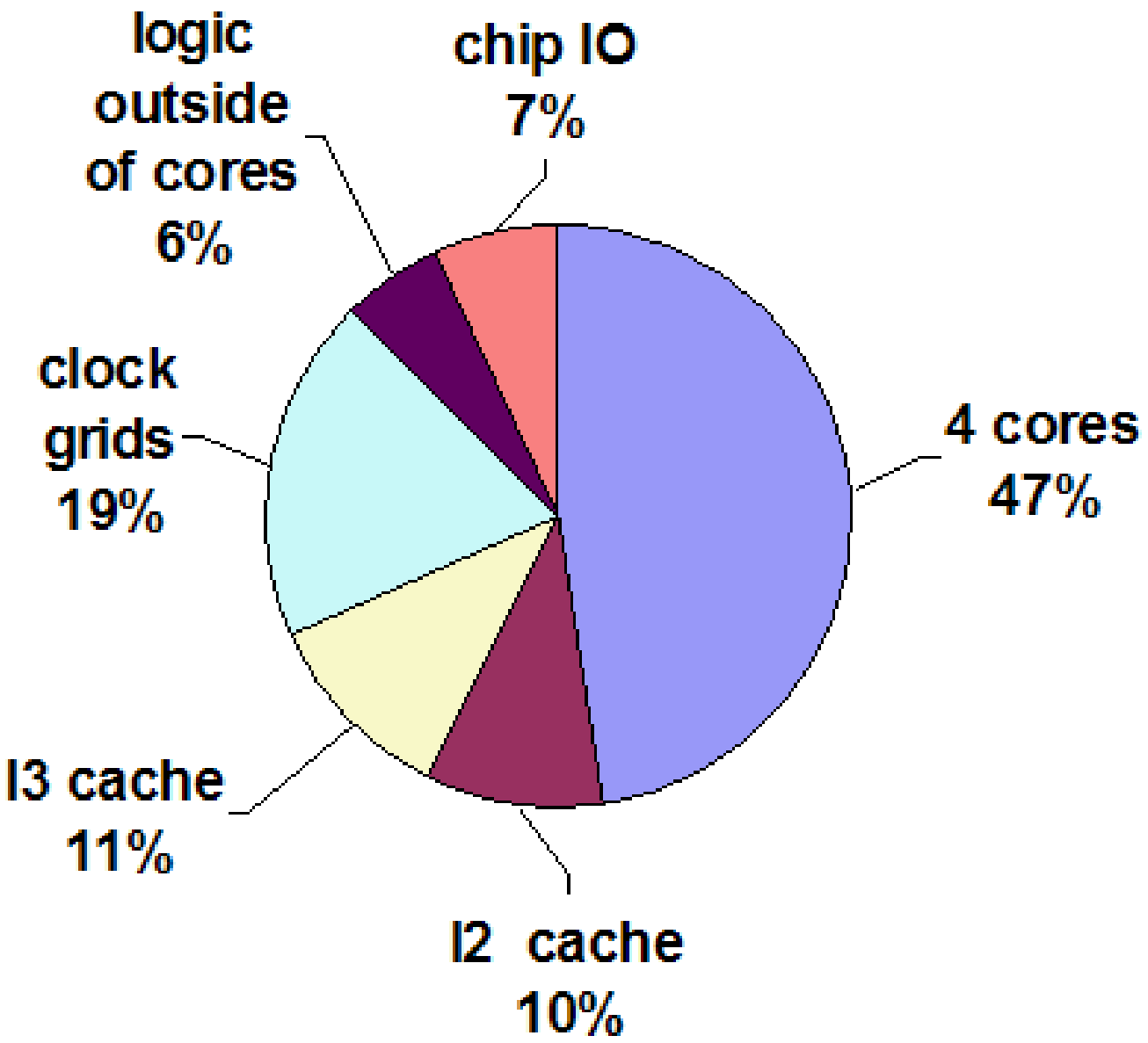
- Fixed chip power budget
  - Roughly same budget as last-generation design. But:
    - Higher frequency
    - Larger chip area
    - Higher capacitance density (from technology scaling)
  - Net: Design team faced significant power issues
- Focused effort on power reduction
  - Keep about same cycle time (in FO4) as previous design
  - But improve power efficiency to enable higher frequency
  - Net: power efficiency improved by ~ 25%
  - Translation: ~ 8-10% improved frequency at const power

Chip Power Methodology

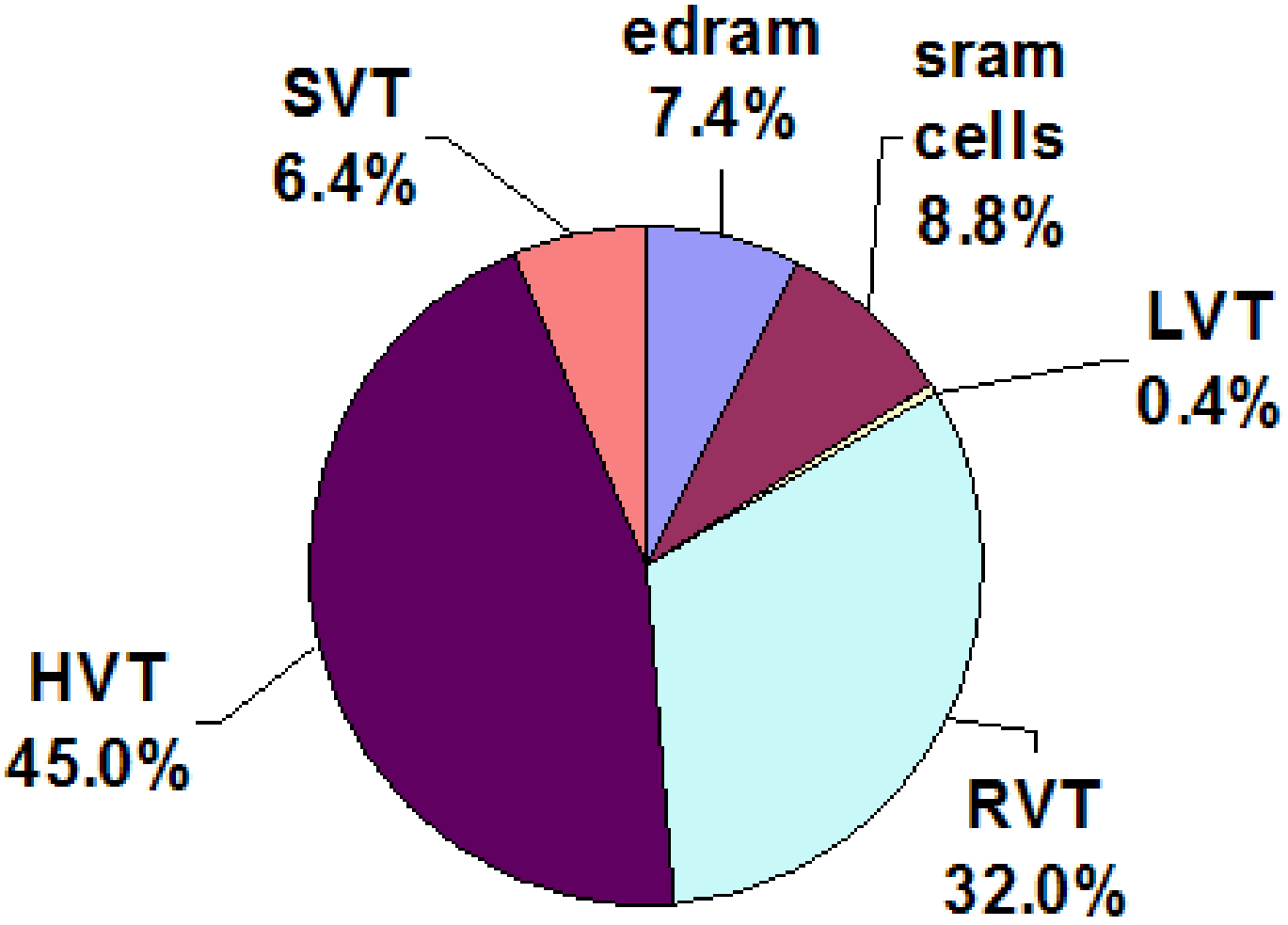


- Power vs input switching factor
- Power vs clock gating percentage
- Leakage Power

Chip Power Breakdown



DC Leakage Breakdown by Device Type



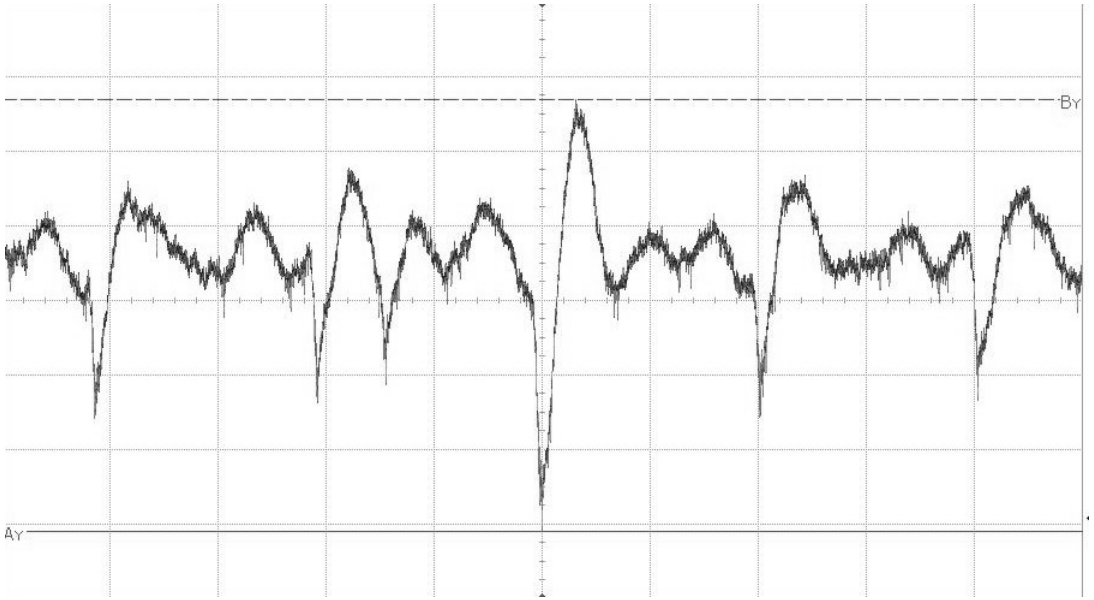
DC leakage = 30% of total power

Power Supply Noise Considerations

- Significant work on power efficiency, clock gating
  - Increases gap from minimum to maximum power states
- Sudden switching current transients, but multi-cycle response time through package
  - Need good amount of on-chip decoupling capacitance
- > 10 μF capacitance added on base power supply
  - Use DRAM trench for dense decoupling cap
- Decoupling also needed on array supplies
  - SRAM, DRAM
  - Early hardware: noise on DRAM supply from bursts of high refresh/access activity

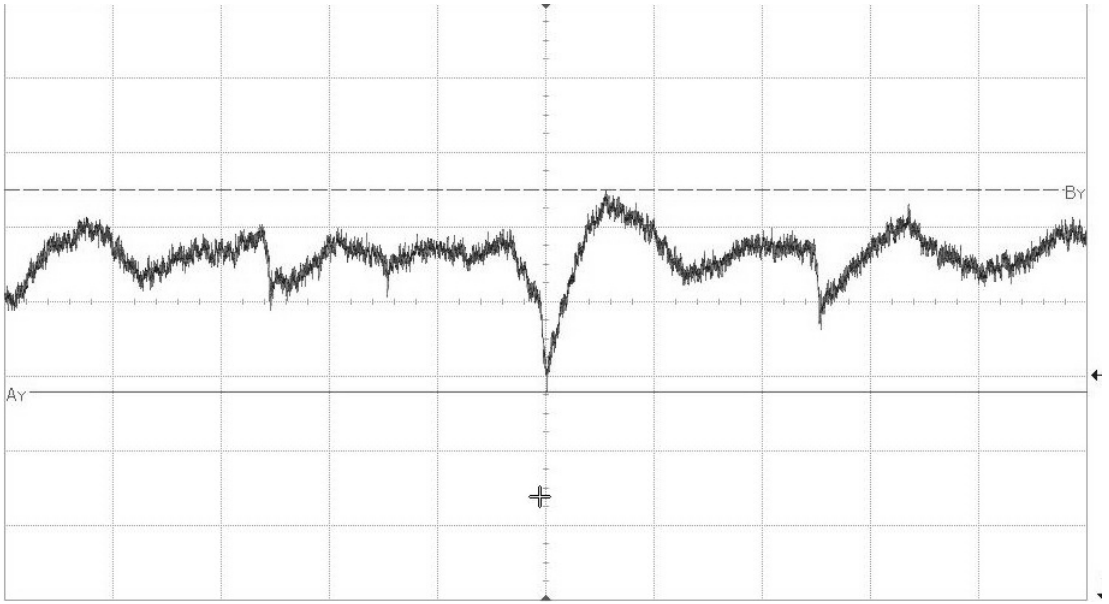
DRAM Supply Voltage Noise

Early hardware, 265nF cap on array supplies



~313mV peak-to-peak

Later hardware, 1.74uF cap on array supplies



~136mV peak-to-peak

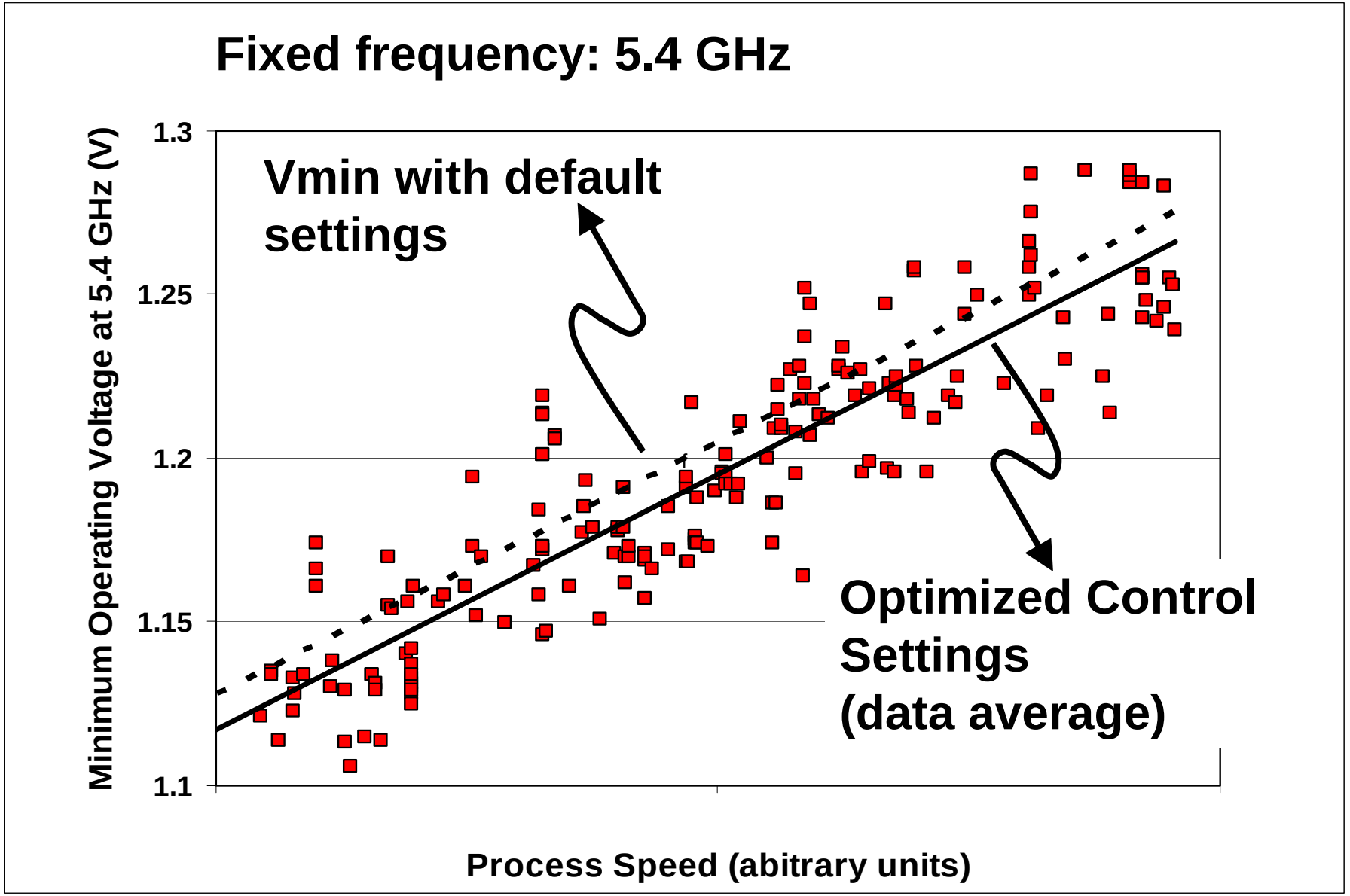
Outline

- Introduction
- Technology and chip overview
  - RAS
  - Clock grids
- Core circuit design
- L3 design
- Power and noise considerations
- Chip frequency tuning
- Conclusions

Hardware Frequency Tuning

- Local (macro-level) controls with fine granularity
  - Local clock pulse width (narrow, nom, wide)
  - Local clock pulse timing (nom, late)
  - Master-clock falling edge delay (MSFF designs)
  - Local clock gating override
  - Array and register files: pulse width & various timing settings
- Global controls: clock duty cycle adjustments
- Other controls for debug and critical path isolation
- Control settings optimized for max overall f_{max}

Chip V_{min} vs Process Speed



Outline

- Introduction
- Technology and chip overview
  - RAS
  - Clock grids
- Core circuit design
- L3 design
- Power and noise considerations
- Chip frequency tuning
- Conclusions

z196 Design: Conclusion

- Design team able to maintain high-frequency pipeline of z10 while adding out-of-order execution
- Large on-chip DRAM L3 for additional performance
  - Deep trench decoupling capacitors provide additional frequency leverage
- Technology features + design for power efficiency =>18% freq boost compared to previous generation
- Net: - 5.2 GHz final product frequency
  - up to 40% perf improvement (single thread legacy workload metric)

Acknowledgements

The authors would like to acknowledge the many contributions from the rest of the System z team, the IBM EDA team, and the IBM Technology Development and Manufacturing teams