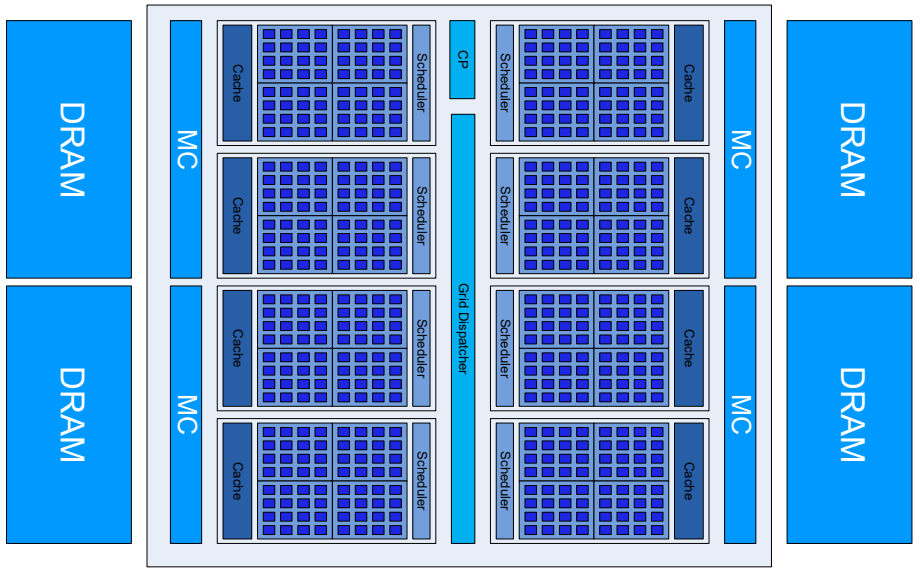
Maximize TOPS (Tera Operations Per Second) per Watt for AI Chip using Early Power Analysis and Reduction

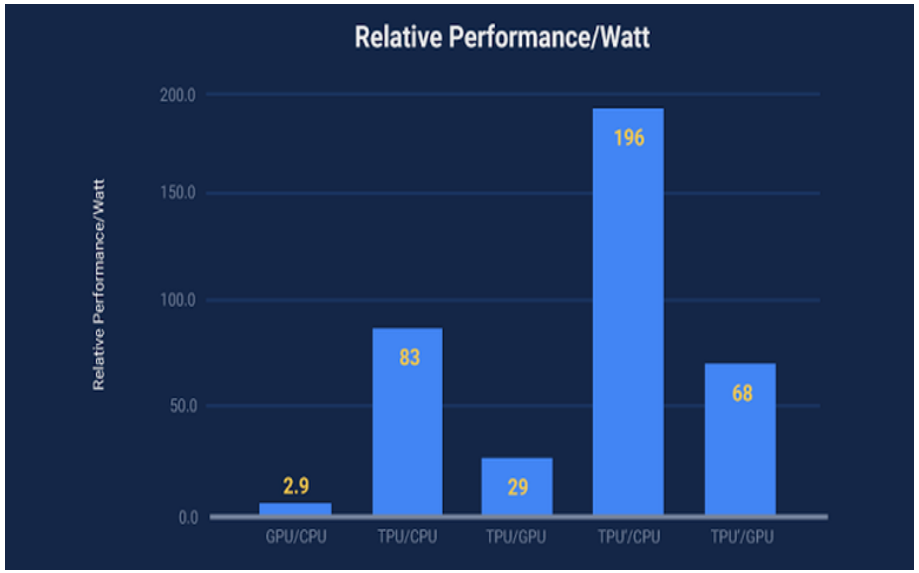
Sun Ling lings@iluvatar.ai
Gan Zhenhua zhgan@iluvatar.ai
Zhang Du du.zhang@iluvatar.ai
Peng Cheng cheng.peng@ansys.com

Motivation for Energy Efficiency for AI Applications

- **Power Efficiency** (TOPS per Watt): One of the most important metrics for Artificial Intelligence ASIC. Reducing power leads to improvement in energy efficiency. So, low-power is critical for AI chip design and its competitiveness in the market. However, netlist power tool cannot provide power reduction suggestion.
- **Quick Design Iteration**: With rapid change of AI algorithm, designers need to analyze power weekly even daily. While the conventional netlist power flow is too late and slow, cannot achieve quick design iteration.



AI ASIC architecture (source: Iluvatar)



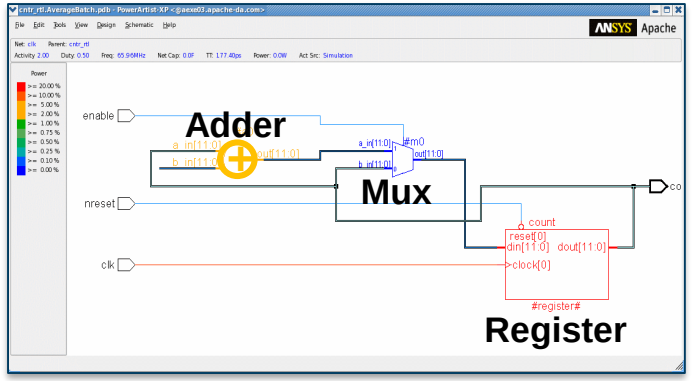
Google TPU energy efficiency (source: Hot Chips 2017)

Early Power Methodology

- **Early RTL Power Reduction and Debug**
  - Identify power reduction opportunities early at RTL using tools and interactive debug, prioritize power reductions, and revise the RTL source code.
  - Use power and power efficiency metrics such as flop and memory clock gating efficiency to track the power efficiency of design, as an RTL sign-off metric.

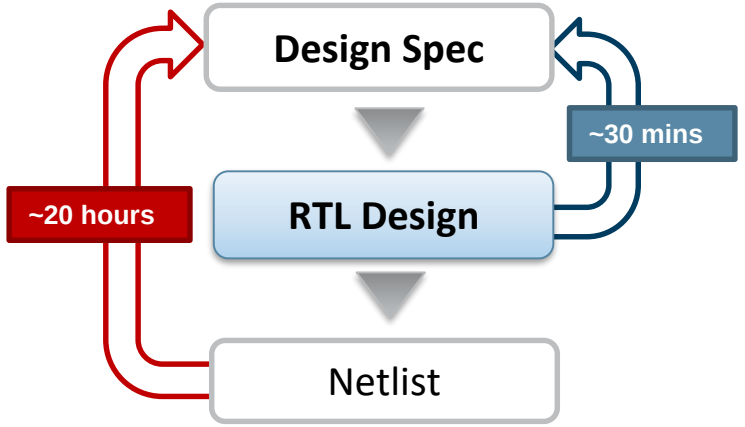
- **Early and Fast Design Iteration**
  - RTL power flow is much faster than conventional netlist power, no need to synthesize RTL into netlist and generate netlist level stimulus. It achieves quick power iteration.

Using visual power debug to identify power hotspots



Tool: ANSYS PowerArtist

Quick design iteration



RTL power analysis

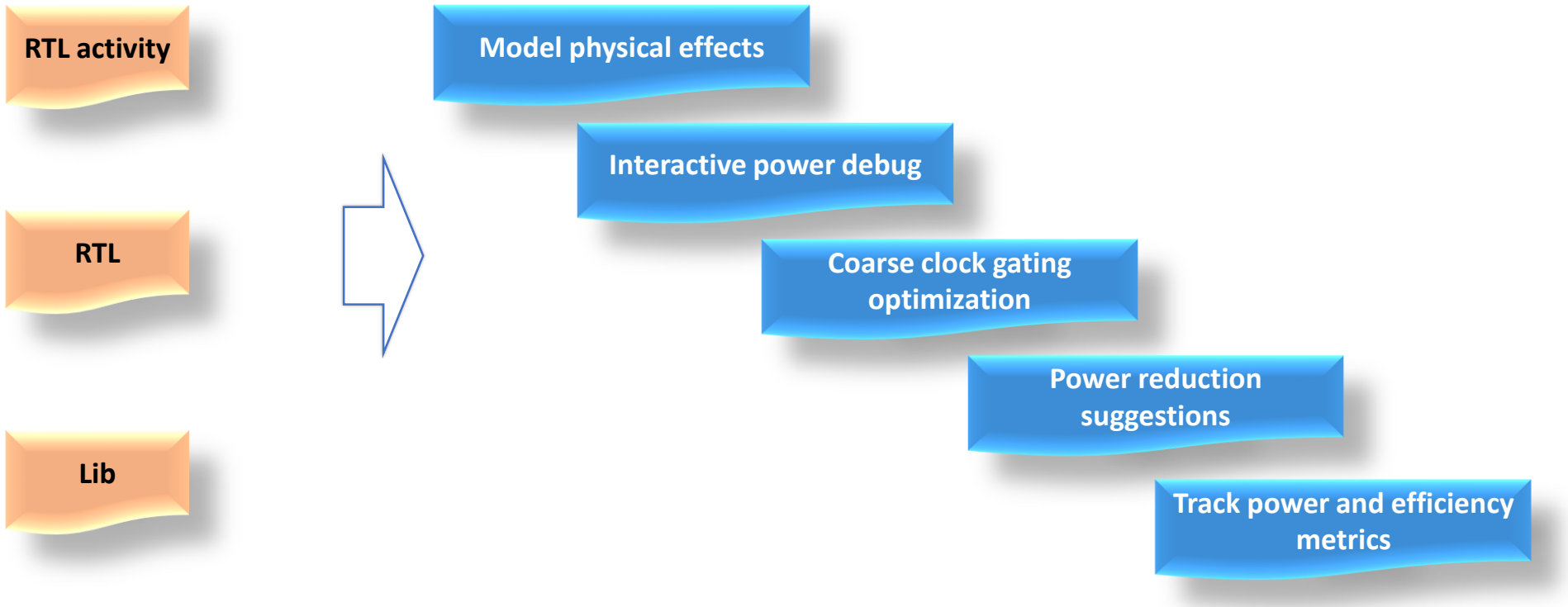
Gate power analysis

RTL Power Analysis and Reduction Core Technology

- Comprehensive early RTL power analysis and reduction methodology includes:
  - Modeling physical effects that are not available in RTL source code: clock distribution network (buffers, clock gates), wire capacitance, cell mapping including multiple threshold libraries. This is key to get predictable power savings.
  - Using an effective and interactive power debug GUI to identify power hotspots, trace activity upstream and downstream to find root cause and to reduce the power-impact. Designers know their RTL best.
  - Using industry tools to identify power reduction opportunities including but not limited to the following: optimize high level clock gating to identify more architectural clock gates, utilize stability and observability of blocks/flops/memory outputs to optimize clock gating enables to shut off more blocks/flops/memories and for more clock cycles, data gating, etc
  - Utilize high capacity and performance for regular power tracking across design iterations

Overview of Early RTL Power Flow

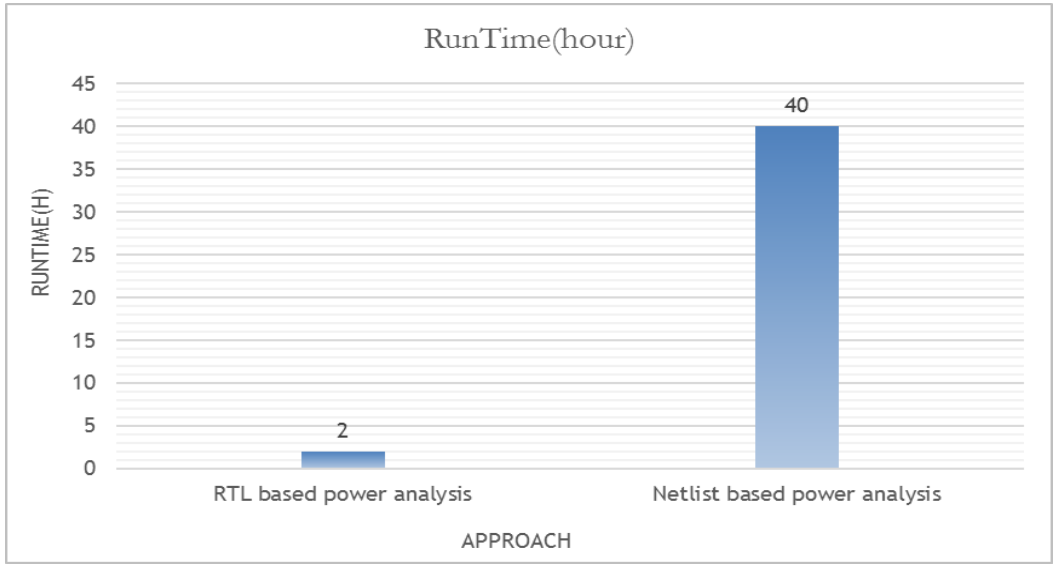
- Overview of early RTL power flow



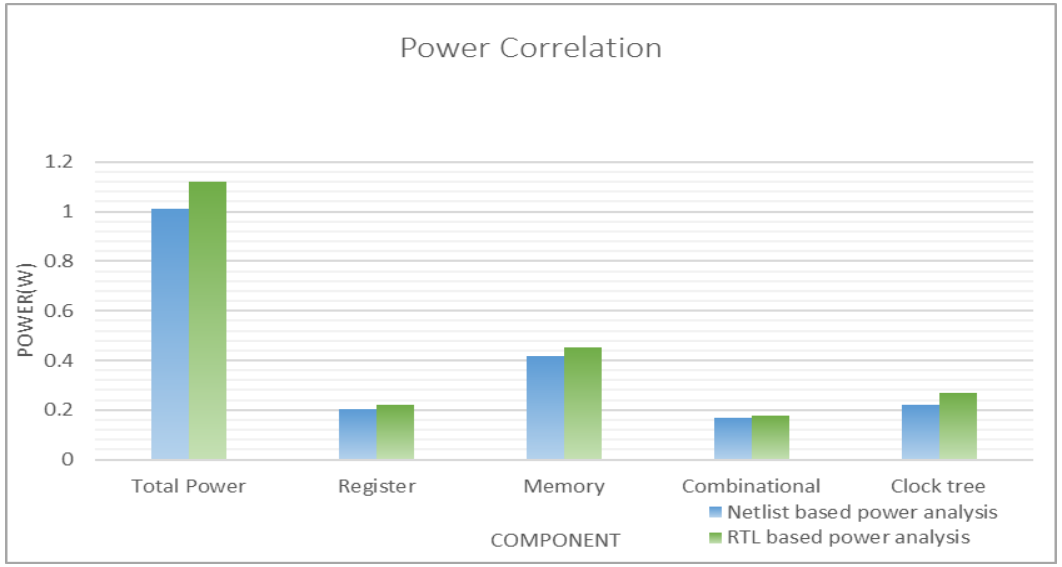
Performance and Accuracy of RTL Power flow

- Performance and Accuracy comparison to netlist power flow

Performance



Accuracy



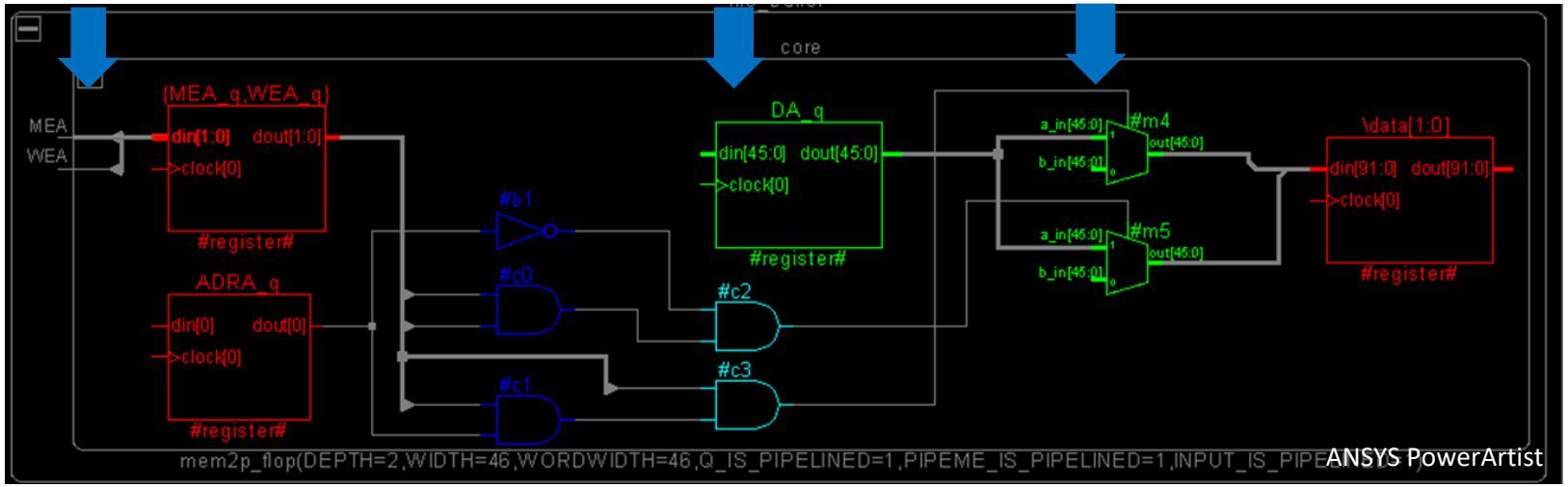
- **20X** faster than netlist power analysis flow

- Within **15%** power difference compared to netlist power

RTL Power Reduction Case on AI Chip

- Observability-don't-care based technique(ODC) saved **5%** power on memory control sub-system, which is instantiated dozens of times in an AI core.

Candidate enable **Candidate register** **Observability condition**



Observability condition signal:

S1=MEA_q&WEA_q&ADRA_q ,

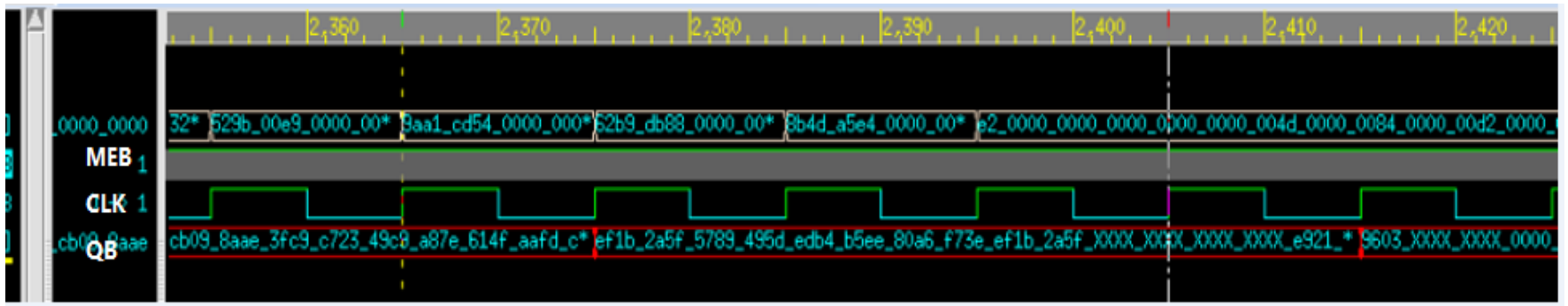
S2=MEA_q&WEA_q&~ADRA_q ;

When S1+S2=0, MEA_q&WEA_q=0, register output is not observed and can be gated.

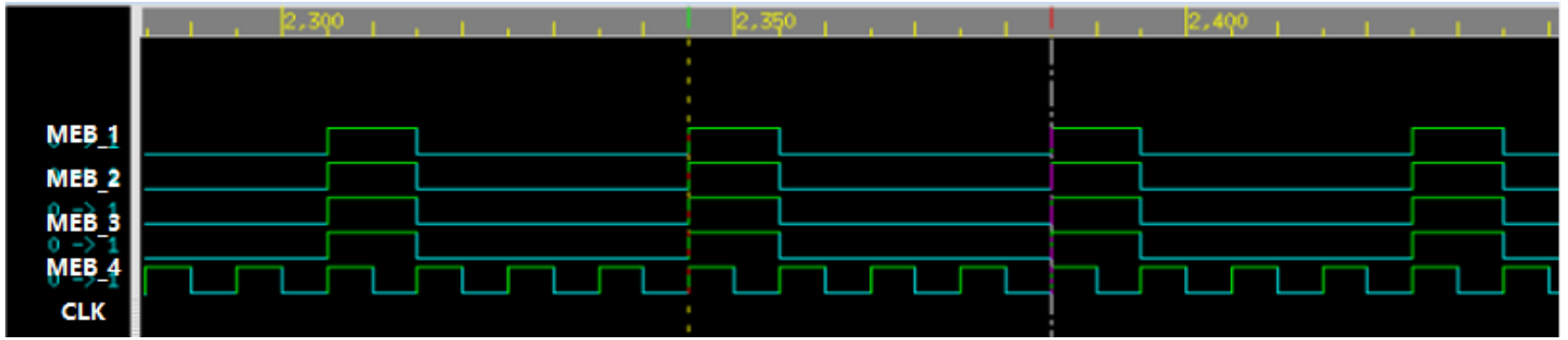
Up-tracing a flip-flop, tool give the candidate CG enable "MEA&WEA" for the candidate register

Memory Power Reduction Case on AI Chip

- Using PowerArtist to identify unreasonable memory power, after optimizing structure, the total power has been reduced from 80W to 40W



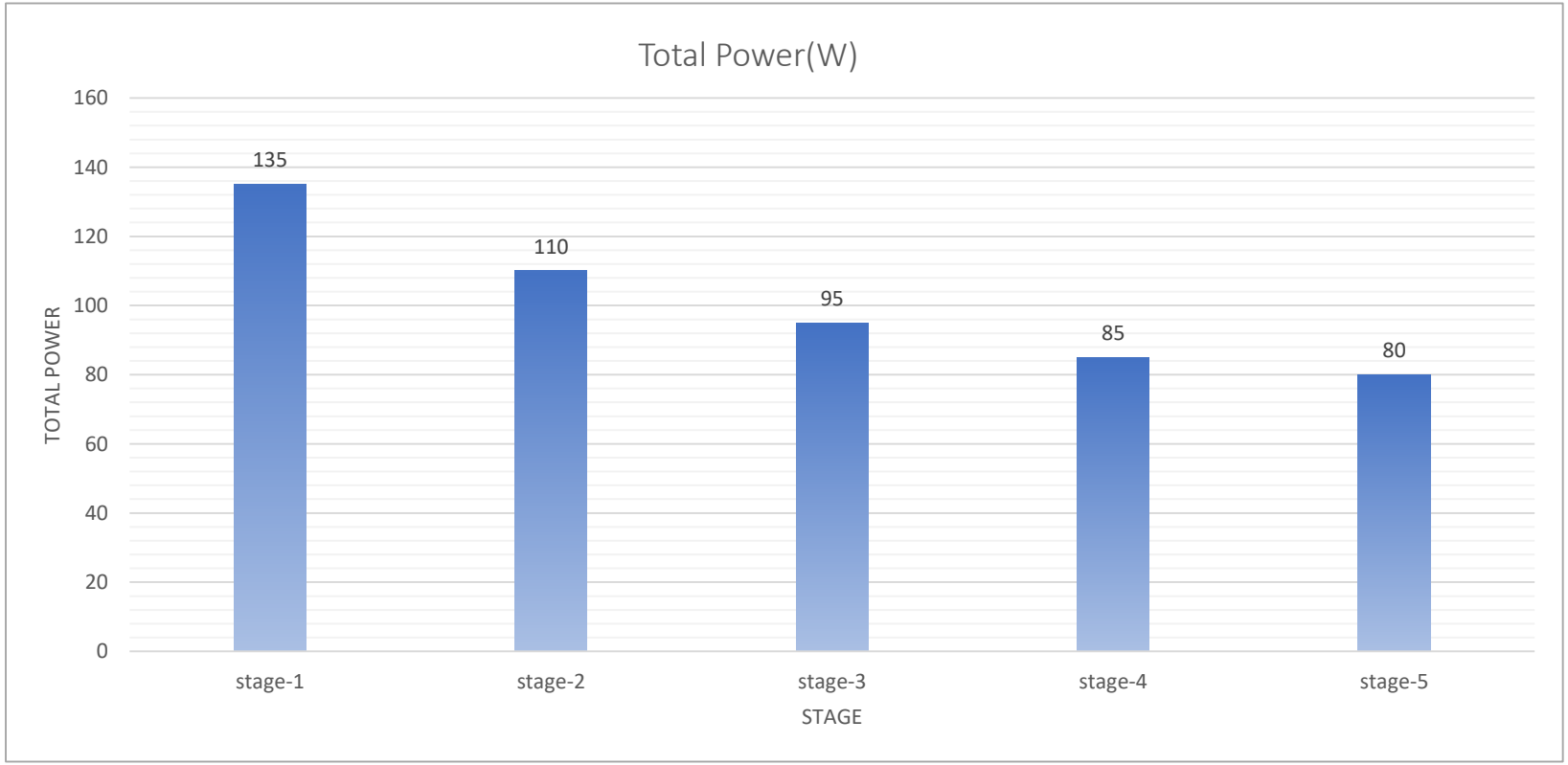
1) Before optimization, memory read in every clock cycles when read enable MEB is always high.



2) After optimization, the original memory is split into 4 sub memories and each only read one clock cycle in four cycles.

Power Regressions

- Monitor power creep across design development cycle
- 40% power reduced (135W to 80W) by RTL power scrubbing in design stage.



Summary

- TOPS per Watt is a critical metric that AI chips compete on.
- Early RTL power analysis flow and methodology enabled reasonable accuracy and much faster TAT than conventional netlist power flow. Designers could monitor power trend during entire design stage earlier and track changes as the AI algorithms are optimized.
- With power debug and reduction techniques of industry-standard RTL power tools, designers can quickly find out power bugs and reduce power earlier, and improve energy efficiency (Tops/W) for artificial intelligence ASIC by reducing power by 40%.