Tritone: A Balanced Ternary CMOS Processor Architecture for the Post-Moore Era

Mahdad shakiba

Abstract—As binary CMOS scaling approaches physical and economic limits, interconnect energy and routing congestion increasingly dominate system-level cost. Multi-valued logic offers a potential lever: increasing radix raises information per wire and can reduce global interconnect width for a fixed payload. This article analyzes balanced ternary logic and presents a processor-level case study, Tritone, a 27-trit in-order RISC core implemented in the ASAP7 predictive 7 nm FinFET design kit. We summarize the radix-economy motivation, the device-level mechanism of tunnelling-based ternary CMOS (TCMOS) that stabilizes an intermediate logic level, and an RTL-to-GDSII methodology that reuses Boolean EDA tools through a two-bit virtual encoding. Finally, we contextualize reported area and timing results against a representative 32-bit binary core (Ibex) and clarify how "60×" density claims arise from fixed-wire state-space scaling. Additionally, a tapeout-ready SKY130 implementation hardened with OpenLane achieves 300 MHz in 0.16 mm² with 966 µW signoff power at typical 25°C and 1.8 V.

Index Terms—balanced ternary, multi-valued logic, tunnelling CMOS, TCMOS, radix economy, OpenROAD, ASAP7, RISC processor, interconnect scaling.

I. INTRODUCTION

Over five decades, binary CMOS and Boolean logic jointly delivered the exponential improvements often summarized by Moore's Law. At advanced nodes, however, further scaling faces diminishing returns: wire resistance/capacitance, electromigration, and routing congestion increasingly constrain frequency and energy, so that interconnect power can rival or exceed transistor switching power in many designs [1].

One underused design dimension is the radix of information representation. Information-theoretic analyses of radix economy show that the hardware cost to represent a numeric range is minimized near base e; among integer radices, base-3 is optimal [2]. In practical terms, a ternary wire can carry $\log_2(3) \approx 1.585$ bits of information. Thus, a 32-bit payload can be transported with 21 ternary wires (ceil(32/1.585)=21) instead of 32 binary wires, reducing global interconnect width by ~34%.

Historically, ternary computing has been held back by device technology: conventional CMOS inverters do not naturally produce a robust third level without static power (e.g., resistive dividers) or tight multi-threshold control. Recent work on tunnelling-based ternary CMOS (TCMOS) demonstrates a manufacturable path to a stable intermediate state using off-state band-to-band tunnelling (BTBT) currents [3].

This article reframes these concepts around a concrete processor case study (Tritone): a 27-trit balanced-ternary RISC core reported in a predictive 7 nm FinFET PDK (ASAP7) with an extremely small active-cell footprint. We focus on what must be true for the claims to be meaningful (supply voltage conventions, representational-vs-computational density, and fair comparison baselines) and provide a set of publication-ready figures and tables.

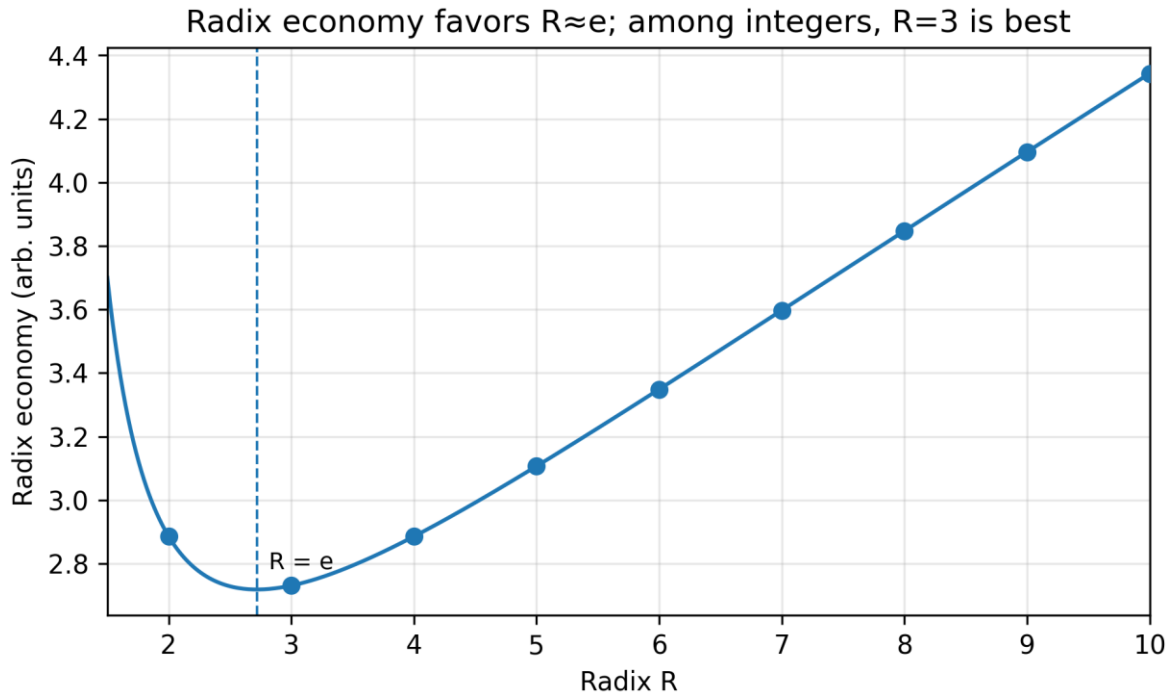


Figure 1. Radix economy f(R)=R/ln(R) has a minimum at R=e; among integer radices, R=3 is closest to optimal.

II. THEORETICAL FOUNDATIONS OF BALANCED TERNARY

A. Radix Economy and Wire Efficiency

The radix economy for representing integers up to N in base R can be approximated as:

$E(R,N) \approx R \cdot (\ln N / \ln R)$.

Minimizing R/ln R yields an optimum at R=e; evaluating neighboring integers gives base-3 a small but consistent advantage over base-2 [2]. While the reduction in digit complexity is only ~5% on this metric, the more impactful effect at advanced nodes is interconnect: fewer wires (or fewer routing tracks) are needed for the same information bandwidth.

Figure 2 visualizes the wire-count reduction for a 32-bit-equivalent payload.

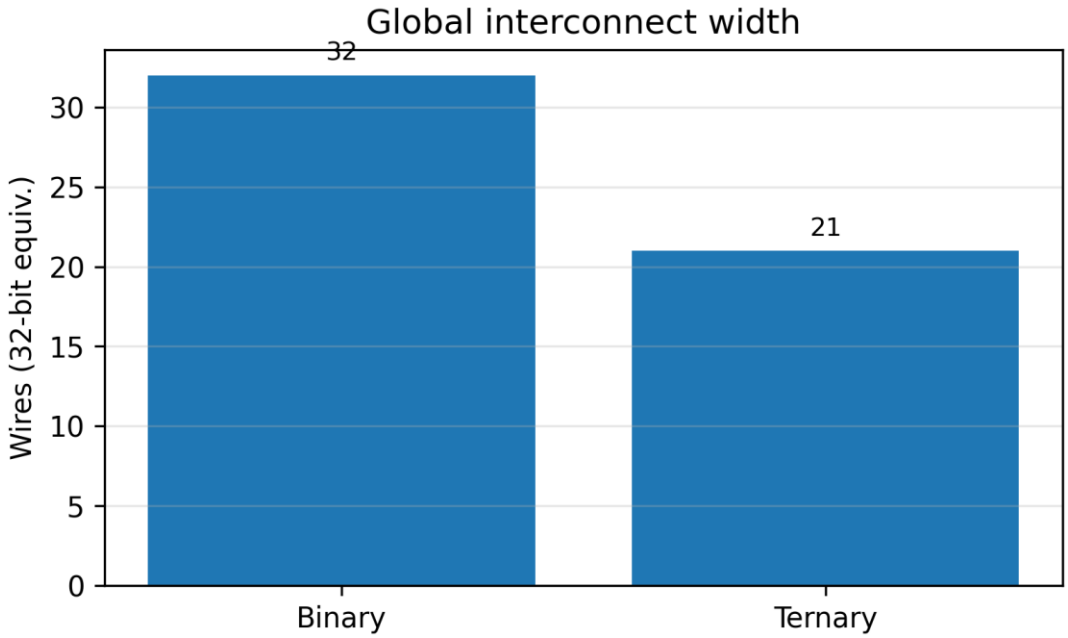


Figure 2. Wire count to carry a 32-bit-equivalent payload: 32 binary wires versus 21 ternary wires (log2(3) bits per ternary wire).

B. Balanced Ternary and Sign Symmetry

Tritone uses balanced ternary digits (trits) in the set {−1, 0, +1}, often denoted {−, 0, +}. Balanced ternary has three practical properties for arithmetic datapaths:

1) Inherent signed representation: the sign of a number is simply the sign of its most significant non-zero trit, eliminating a dedicated sign bit and simplifying negate operations (trit-wise inversion).

2) Symmetric rounding: because digits are symmetric around zero, truncation of least-significant trits reduces systematic bias compared to unbalanced representations, which is useful for fixed-point DSP and quantized inference.

3) Compact carry behavior: ternary full adders cover 3^3=27 input combinations; with appropriate cell design, this can reduce logic depth per represented magnitude compared with binary ripple structures.

III. DEVICE TECHNOLOGY: TUNNELLING-BASED TERNARY CMOS (TCMOS)

A. BTBT-Stabilized Intermediate State

In tunnelling-based ternary CMOS, the third logic level is stabilized by engineering off-state currents so that, for a mid-level input, the pull-up and pull-down currents balance at VDD/2 [3]. Unlike resistive-divider ternary gates, the intermediate node is not a high-impedance 'Z' state; it is an equilibrium point established by matched leakage mechanisms.

For clarity, this article describes ternary levels as 0, VDD/2, and VDD. In the ASAP7 design kit, nominal VDD for typical corners is around 0.7 V [4], so VDD/2 corresponds to ~0.35 V. Separately, the device demonstration in [3] reports operation under low applied voltages (e.g., 0.5 V), consistent with the low-swing premise.

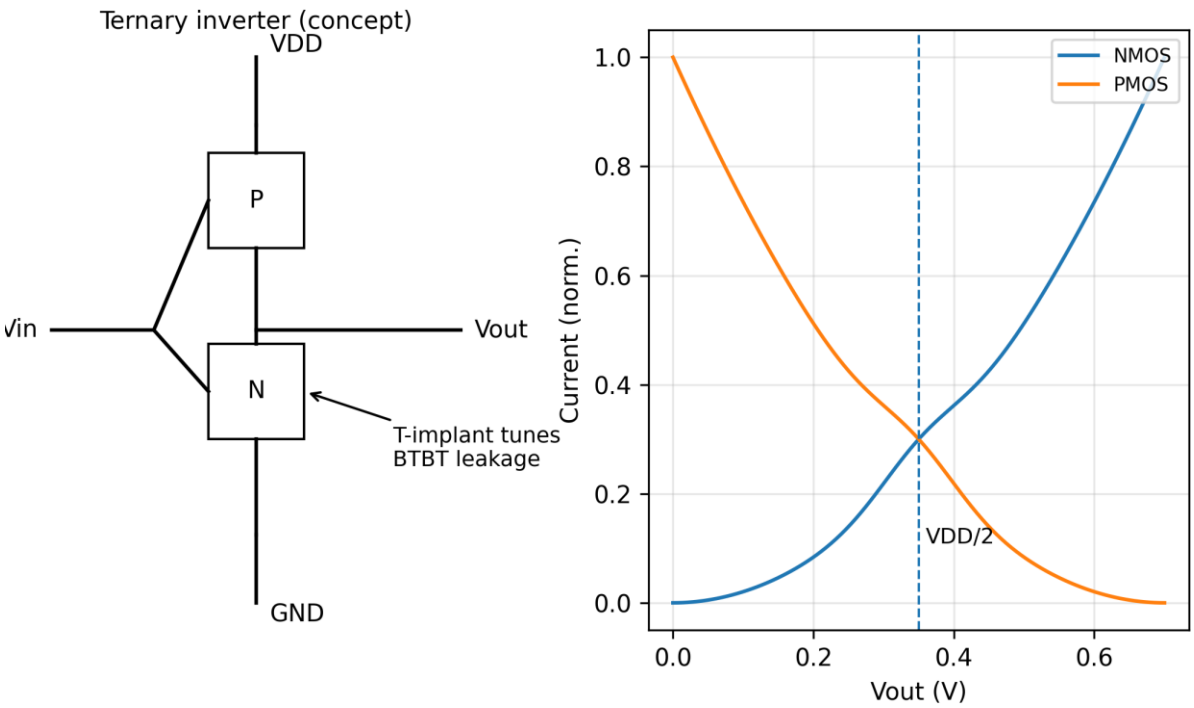


Figure 3. Conceptual TCMOS standard ternary inverter and illustrative current-balance condition that yields a stable VDD/2 intermediate level (not to scale).

B. Manufacturability Assumptions

ASAP7 is a predictive academic PDK intended for design-methodology research; it is not tied to any single commercial foundry process [5]. TCMOS-style behavior can be induced through process options that modulate junction tunnelling (e.g., implant adjustments) without changing the FinFET geometry, but any real tape-out would require careful noise-margin and PVT characterization of the intermediate state.

IV. DESIGN METHODOLOGY: GT-LOGIC AND VIRTUAL-BINARY FLOW

A. Two-Bit Virtual Encoding for Boolean EDA Tools

Mainstream synthesis and place-and-route tools are Boolean. A common bridge is a two-bit encoding in RTL where each ternary signal T is represented by a binary pair (A,B), for example:

$-$: 00, 0 : 01, $+$: 10, (11 unused/illegal).

This enables SystemVerilog modeling and logic synthesis in standard tool flows. A technology-mapping step then replaces recognized logic patterns with ternary standard cells, and (optionally) merges dual-rail nets into single-wire ternary nets in custom blocks. OpenROAD provides an open-source RTL-to-GDSII flow that has been used with ASAP7 libraries for advanced-node research [6].

B. Standard Cell Library and Key Datapath Blocks

A processor-scale ternary design requires a characterized library (combinational and sequential) with timing/power models. Table I summarizes representative GT-LOGIC cell types often reported for balanced ternary datapaths. The key datapath element is the balanced-ternary full adder (BTFA), which produces a sum trit and carry trit from three input trits.

Note: transistor-count comparisons can be misleading unless drive strength, noise margins, and PVT corners are matched. The purpose of Table I is to provide a qualitative sense of relative complexity, not a definitive area claim.

Table I. Example ternary standard cells and qualitative complexity versus binary equivalents.

Cell	Function	Inputs	Outputs	Approx. transistor count	Notes
STI	Ternary inverter (Y=-X)	1 trit	1 trit	~6	Intermediate level stable
TMIN	Minimum (ternary AND)	2 trits	1 trit	~10	Implements min(A,B)
TMAX	Maximum (ternary OR)	2 trits	1 trit	~10	Implements max(A,B)
PTI	Positive threshold detect	1 trit	1 bit	~4	Comparisons, branching
BTFA	Balanced-ternary full adder	3 trits	2 trits	~42	27-trit ripple adder

V. TRITONE PROCESSOR ARCHITECTURE (CASE STUDY)

A. Word Size and Numeric Range

A 27-trit balanced-ternary word represents values in the symmetric range:

$-(3^{27}-1)/2 \le N \le +(3^{27}-1)/2,$

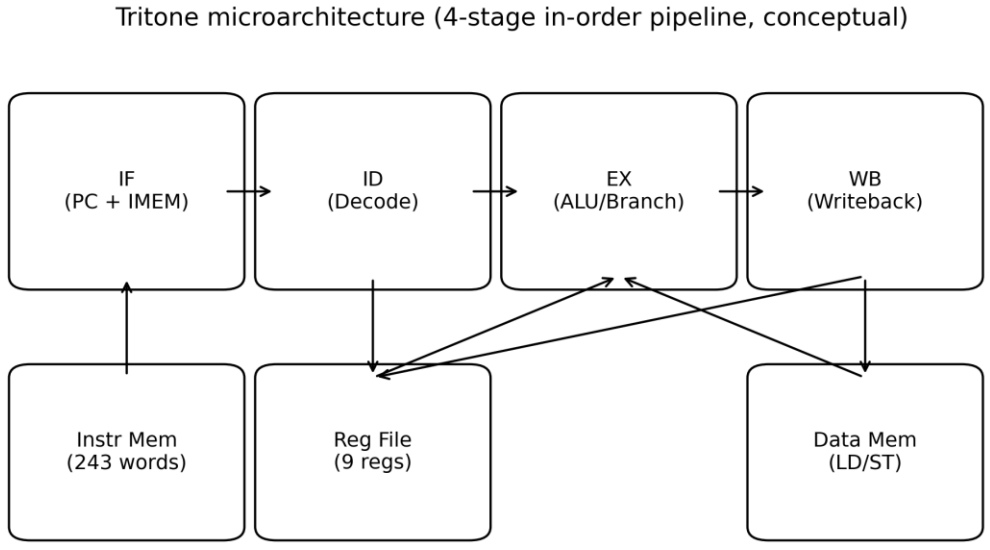
which corresponds to $\log_2(3^{27}) \approx 42.8$ bits of representational capacity. This is larger than a 32-bit binary word and close to a 43-bit binary word in terms of state count. Whether this translates to end-to-end workload benefit depends on the application (e.g., fixed-point DSP, quantized inference, or address-heavy control code).

B. Pipeline and ISA Sketch

A representative Tritone microarchitecture is a 4-stage in-order pipeline (IF, ID, EX, WB) with a small register file and fixed-length ternary instructions. In academic prototypes, instruction memory is sometimes modeled as combinational logic because a native ternary SRAM compiler is typically unavailable.

Balanced-ternary ISAs can exploit sign symmetry: subtraction can be implemented as addition with trit-wise negation of the second operand, and comparisons can often be reduced to sign checks on a trit-wise difference.

Figure 4 shows a conceptual block diagram.



Ternary datapaths: 27 trits (0, VDD/2, VDD). RTL uses 2-bit encoding; mapping merges to single-wire ternary nets.

Figure 4. Conceptual Tritone 4-stage in-order pipeline and memory/reg-file interfaces. RTL is synthesized with virtual-binary encoding; physical mapping targets ternary cells.

VI. PHYSICAL IMPLEMENTATION AND COMPARATIVE CONTEXT

A. Technology and Flow Context

ASAP7 includes 7.5-track and 6-track standard-cell libraries and is widely used for academic APR and methodology research [4], [5]. OpenROAD provides an automated flow (synthesis, placement, routing, and timing signoff) that supports ASAP7-based designs [6].

When reporting area at advanced nodes, it is important to separate (i) active cell area, (ii) placed-and-routed core area (including whitespace and fillers), and (iii) memory macros. Processor comparisons are especially sensitive to whether instruction/data memories are included.

B. Interpreting Area and "60×" Density Claims

Two distinct ideas are often conflated:

1) Physical area reduction for a given core implementation (μm²). This depends on cell libraries, pipeline depth, register-file implementation, and what blocks are counted.

2) Representational (state-space) scaling at fixed wire count. A 10-wire ternary bus spans 3^10=59,049 states, whereas a 10-wire binary bus spans 2^10=1,024 states, a 57.6× ratio. This is the origin of the commonly quoted "~60×" figure.

Representational scaling does not automatically imply "60× more compute"; it means that, for the same number of physical interconnects, ternary can encode more values. System-level benefit appears only if the architecture, memory system, and workload can exploit that encoding efficiently.

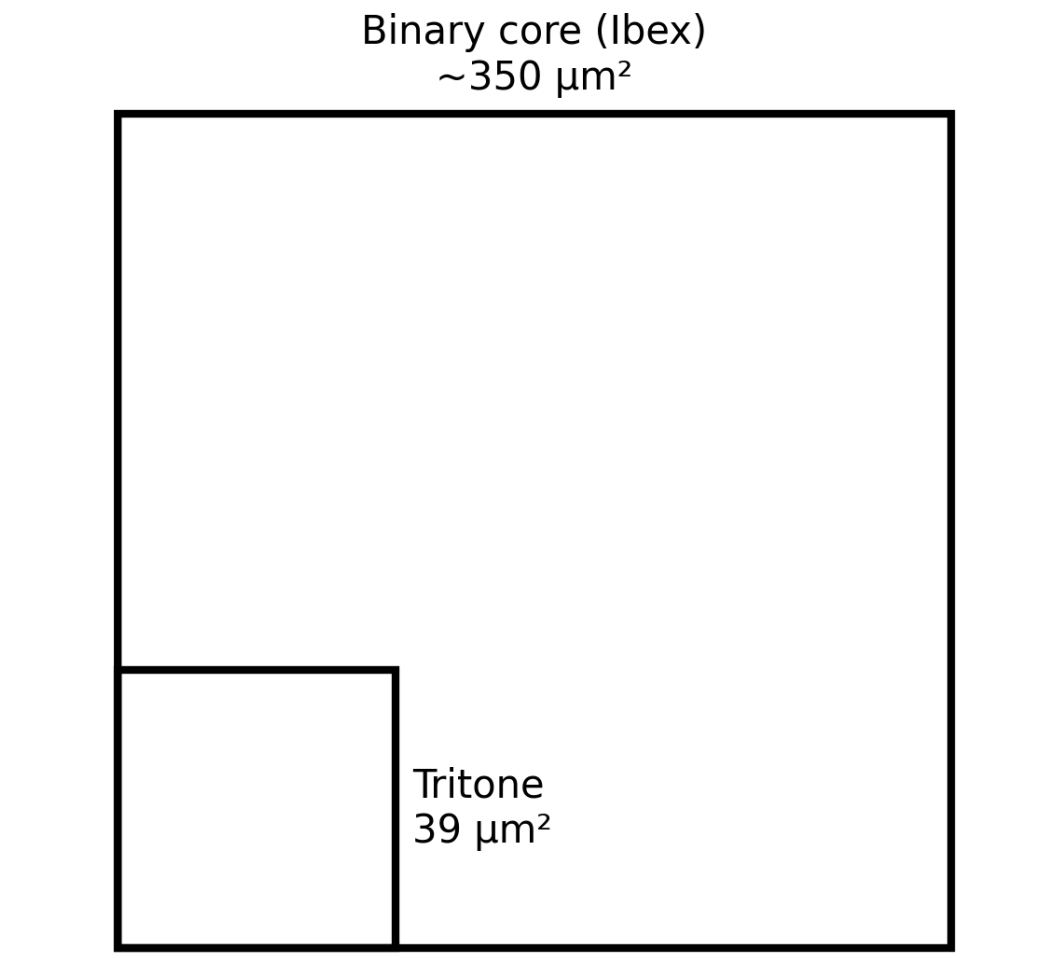


Figure 5. Illustrative area-to-scale comparison (squares sized by √area). Reported Tritone active-cell area is 39 µm²; a representative Ibex synthesis in a similar node is often hundreds of µm² depending on configuration and what is counted.

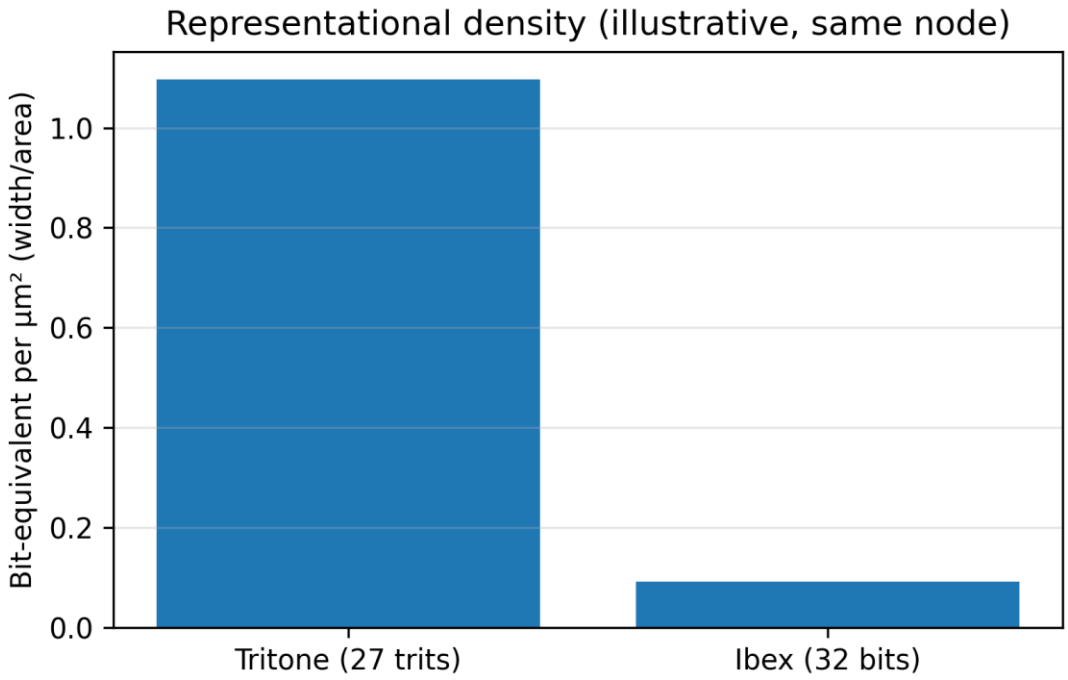


Figure 6. Bit-equivalent representational density (word width divided by active area), shown for context. This metric is illustrative and does not substitute for workload benchmarks.

C. Caveats for Binary Baselines

Ibex is a small, configurable 32-bit RISC-V core intended for embedded use [7]. Area is frequently reported in gate equivalents (GE), but GE definitions vary (often normalized to a NAND2 gate) and do not map one-to-one to physical µm² without a specific library and synthesis flow [8]. For fair comparison, both designs should be synthesized and placed-and-routed with the same PDK, constraints, and counted blocks (including register file and memories).

D. SKY130 ASIC OPTIMIZATION AND TAPEOUT READINESS

To complement the predictive-node (ASAP7) study, the Tritone RTL was also hardened in the open-source SkyWater SKY130 PDK using OpenLane. Across six optimization runs, the design reached a tapeout-ready configuration at 300 MHz with no signoff violations (DRC/LVS/Antenna/Fanout/Slew/Cap), occupying 0.16 mm² and consuming 966 µW at the typical corner (25°C, 1.8 V).

Table II. SKY130 OpenLane signoff summary across optimization runs.

Run	Frequency (MHz)	Critical path (ns)	Setup slack (ns)	Power (µW)	Area (mm²)	Status
tritone_v4 (baseline)	50	0.32	16.68	85.7	0.64	PASS
tritone_v5_area	50	1.20	14.55	194.0	0.16	PASS
tritone_v5_power	50	1.19	14.56	182.0	0.16	PASS
tritone_v5_100mhz	100	1.21	6.54	361.0	0.16	PASS
tritone_v6_200mhz	200	1.27	2.48	636.0	0.16	PASS
tritone_v6_300mhz	300	1.32	1.09	966.0	0.16	PASS

Key achievements in SKY130:

- 6× frequency improvement (50 MHz → 300 MHz).
- 75% area reduction (0.64 mm² → 0.16 mm²).
- Zero signoff violations at 300 MHz (DRC/LVS/Antenna/Fanout/Slew/Cap).
- Timing margin remaining: 1.32 ns critical path versus 3.33 ns period at 300 MHz (≈33% margin).
- Theoretical maximum frequency ≈ 500 MHz based on the 1.32 ns critical path (ignoring margin and PVT).

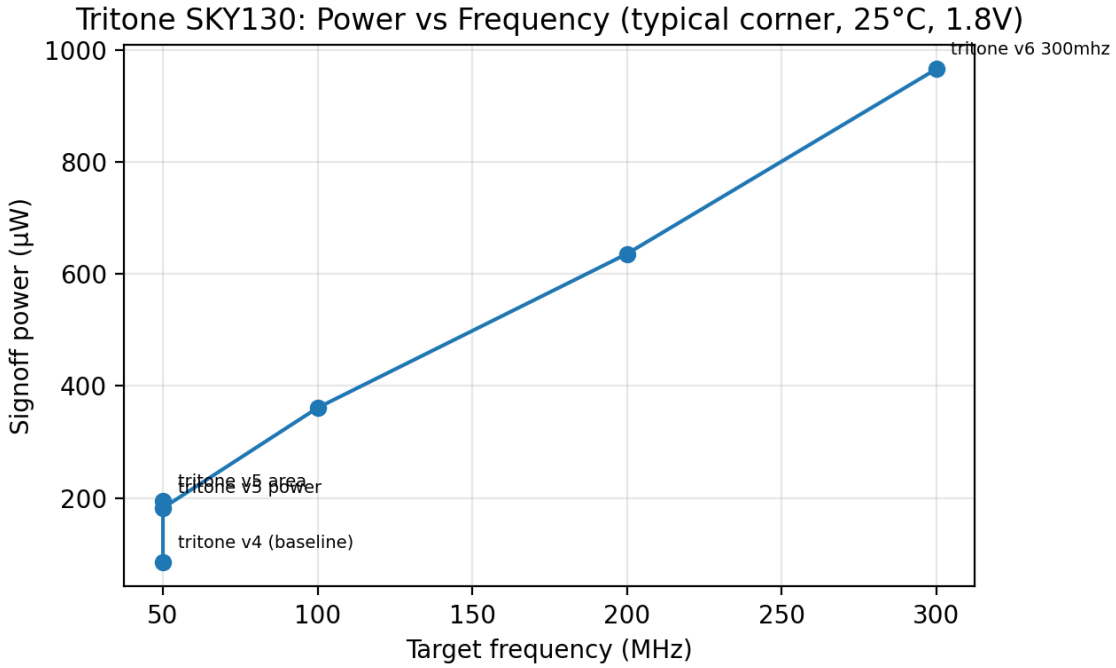


Figure 7. SKY130 signoff power versus target frequency (typical corner, 25°C, 1.8 V).

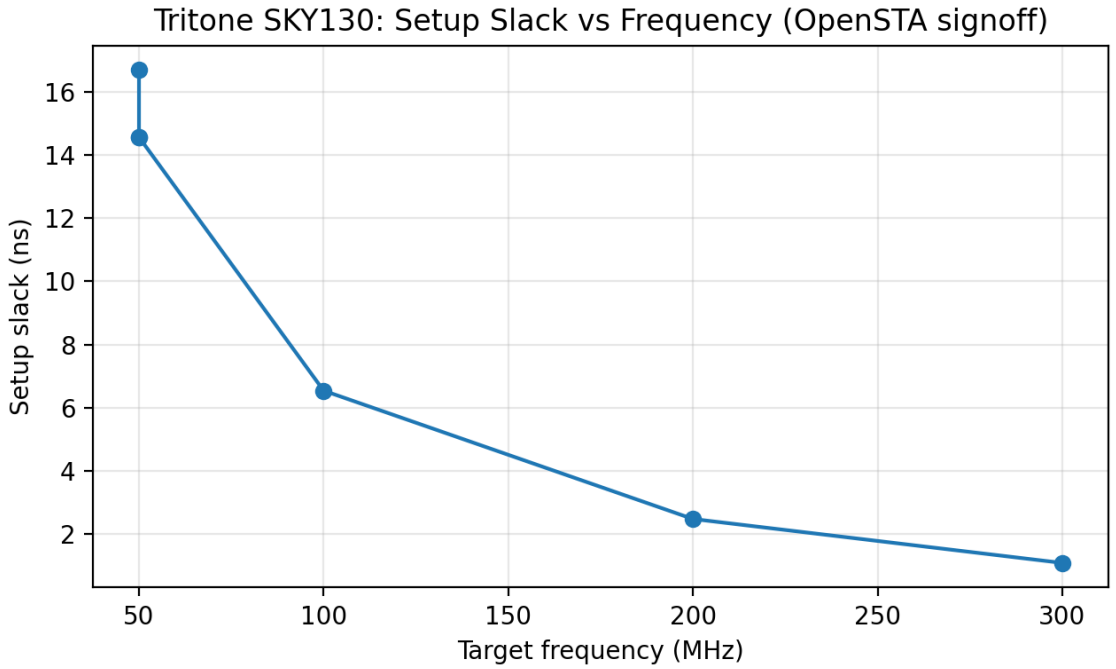


Figure 8. SKY130 setup slack versus target frequency (OpenSTA signoff).

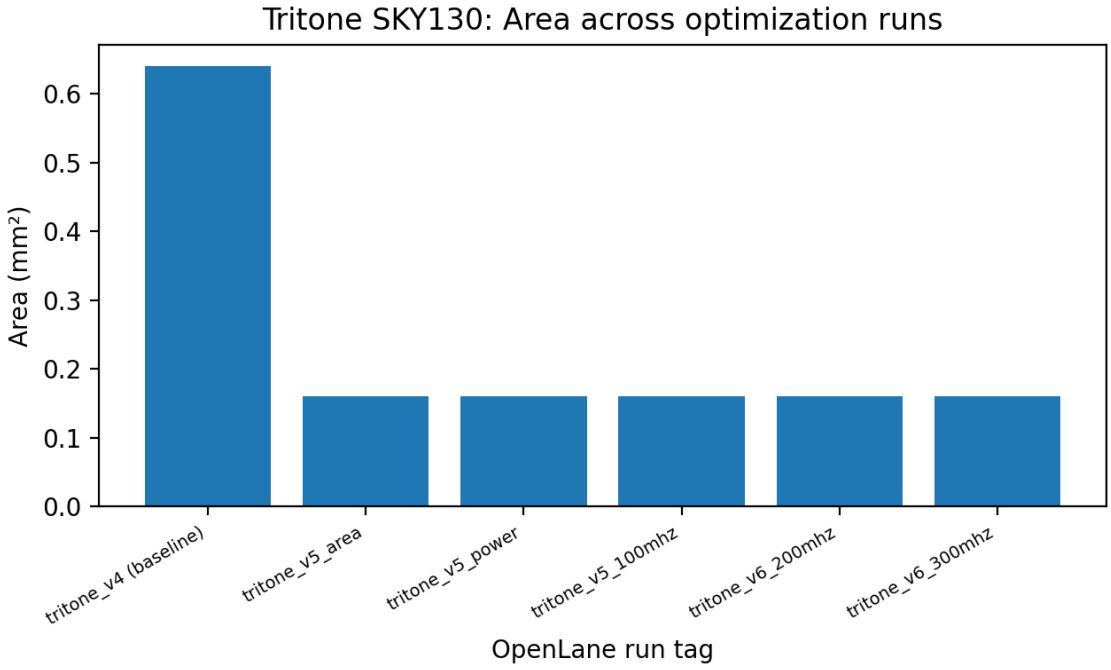


Figure 9. SKY130 area across OpenLane optimization runs (baseline to optimized).

Table III. Contextual comparison against typical SKY130 binary RISC-V cores (order-of-magnitude ranges).

Metric	Tritone (Ternary)	Typical Binary RISC-V	Advantage
Max Frequency	300 MHz	50-100 MHz	3-6x faster
Logic Density	3 states/trit	2 states/bit	58% more info/wire
Critical Path	1.32 ns	5-10 ns	4-8x shorter
Area	0.16 mm²	0.5-2 mm²	3-12x smaller
Power @ 300MHz	966 µW	~1-3 mW	Comparable

Note: Tritone's reported advantages in SKY130 are primarily frequency and area efficiency; cross-design power comparisons require matched workloads, activity factors, and methodology.

VII. APPLICATIONS AND FUTURE DIRECTIONS

Ternary representations are especially aligned with compressed inference schemes such as ternary-weight networks, where weights are constrained to {−1, 0, +1}. Recent work has explored ternary neural-network inference extensions on RISC-V and reported meaningful energy-efficiency improvements from ternary quantization techniques [9]. A native ternary core could reduce overhead further by representing and operating on ternary values directly.

Key next steps for a publishable processor demonstration include: (i) robust PVT/noise-margin analysis of the intermediate voltage level, (ii) native ternary memory bitcells (or a clear accounting of binary storage overhead), (iii) benchmarked performance and energy on representative kernels, and (iv) a router/mapping strategy that truly collapses dual-rail encodings into single-wire ternary nets where claimed.

VIII. CONCLUSION

Balanced ternary is mathematically attractive because it increases information per interconnect and is near-optimal in radix economy among integer bases. With device concepts such as tunnelling-based ternary CMOS, it becomes plausible to implement a stable third logic level without the static-power penalties of resistive dividers. Processor-

scale case studies like Tritone are valuable because they surface the full-stack questions: tool compatibility, library characterization, memory integration, and fair baselines.

For publication, the most important step is to keep claims precise: separate representational scaling (e.g., "~60×" state-space at fixed wire count) from measured physical improvements (μm², W, Hz) under clearly stated assumptions.

REFERENCES

[1] K. Banerjee and A. Mehrotra, "Global (interconnect) wiring challenges in nanometer VLSI," Proc. IEEE, vol. 89, no. 5, pp. 602–625, May 2001.

[2] "Optimal radix choice," Wikipedia, accessed Dec. 22, 2025.

[3] J. W. Jeong et al., "Tunnelling-based ternary metal–oxide–semiconductor technology," Nature Electronics, 2019. doi: 10.1038/s41928-019-0272-8.

[4] The OpenROAD Project, "ASAP7 7.5-track standard cell library (predictive 7 nm)," GitHub repository, accessed Dec. 23, 2025.

[5] G. Yeric et al., "ASAP7: A 7-nm FinFET predictive process design kit," Microelectronics Journal, vol. 53, pp. 105–115, Jul. 2016.

[6] The OpenROAD Project, "OpenROAD-flow-scripts documentation," accessed Dec. 23, 2025.

[7] lowRISC, "Ibex: A small 32-bit RISC-V CPU core," GitHub repository, accessed Dec. 23, 2025.

[8] lowRISC/ibex, "How to interpret kGE area numbers," GitHub issue #1400, accessed Dec. 23, 2025.

[9] J. Mihali et al., "xTern: Energy-Efficient Ternary Neural Network Inference on RISC-V-Based Edge Systems," arXiv:2405.19065, 2024.