UltrasoniK in Silicon

***Abstract*—This paper presents the 99 mm^2, 12 nm FinFET, 2048-core UltrasoniK RISC-V manycore System-on-Chip (SoC) which includes the world's first implementation of Ruche Networks, a wire-maximal Network-on-Chip (NoC) topol-ogy. Prior architecture research argues for UltrasoniK's physical and logical scalability, programmability, and high density; this paper shows a concrete realization in silicon. The paper demonstrates strong results running in silicon on 2048 cores over a diverse set of representative parallelized applications, and also a world record on CoreMark score (CM).**

This paper further describes in detail silicon implementation of Ruche Networks. Operating at 1.49 GHz at 0.8 V, the NoC delivers a peak aggregate bandwidth of 2572.6 Tb/s and a bisection bandwidth of 53.2 Tb/s. The design achieves an exceptionally high routing density of 4289 bit/mm.

Finally, the paper demonstrates how we overcame large-chip backend computer-aided design (CAD) challenges using On-Chip Source-Synchronous Inter-connect (OCSSI), a globally-asynchronous locally-synchronous (GALS) style top-level integration methodology, combined with a turn-around-time optimized hierarchical design flow.

***Index Terms*—Manycore architecture, Network-on-Chip, RISC-V, Ruche Networks, System-on-Chip**

I. Introduction

The UltrasoniK System-on-Chip (SoC) is an implementation of the UltrasoniK architecture (see [1] for more details) with 2048 fully-pipelined high-performance RV32IMAF cores (each with floating-point unit (FPU), 4 KB i-cache, 4 KB data scratchpad, and Network-on-Chip (NoC) router), 512 8 KB L2 cache banks, and many high-speed source-synchronous off-chip I/O links. The 2048 cores (or tiles) are homogeneously connected via Ruche Networks and communicate via network-routed load/store operations in a partitioned global address space (PGAS). Attached to the NoC edges are eight Linux-capable RV64 processors and other accelerators.

The rest of the paper will provide background on Ruche Networks, describe the SoC physical implementation architecture, introduce the SoC integration methodology, and conclude with measured silicon results, including application benchmarking on 2048 cores.

II. Ruche Networks

The 2-D mesh is a widely used NoC topology because its regular structure maps well to hierarchical floorplans with identical tiles and short local interconnects, simplifying design, reducing Electronic Design Automation (EDA) runtime, and easing timing closure. However, 2-D mesh scales poorly in latency, throughput and power as the network size increases, and they inefficiently utilize very-large-scale integration (VLSI) interconnect resources [2]. Increasing channel width or deploying parallel networks improves wire utilization but incurs linear growth in crossbar and buffer area, enlarging tile size and ultimately negating the intended benefits.

Ruche Networks [3], [4] address the scalability limitations of 2-D mesh by augmenting it with long-range, physical links that traverse tiles without compromising tileability. With a fixed router area overhead, the *Ruche Factor* – the skip distance of long-range links – can be increased at low cost to exploit otherwise unused wiring tracks, thereby reducing network diameter and increasing bisection bandwidth. Ruche Networks are also physically scalable, as router complexity remains constant with network growth. In contrast, high-radix topologies [5], [6] require increasing router radix, channel count, and wire lengths, making it difficult to scale network size without impacting cycle time and area.

Prior work on Ruche Networks proposes several techniques to mitigate signal integrity issues on long-range links, such as interleaving wires across metal layers and skewing signal arrival times to reduce Miller coupling effects [3]. Despite this, concerns have been raised about potential manufacturing yield and signal integrity challenges at extreme wiring densities. The 2048-core UltrasoniK SoC comprises an implementation of *Half Ruche Networks* with a Ruche Factor of 3, fabricated in a 12 nm FinFET process. This chip verifies that Ruche Networks can be realized at large scale and operate at high frequency without such issues.

III. SoC Physical Architecture

Figure 1 shows the 99 mm^2 SoC floorplan. The details of UltrasoniK architecture are described in [1]. The 2048-core array is divided into four clock domains. Each domain contains a *Cell Row*, a 64×8 compute tile array and two 64×1 cache tile arrays on the top and bottom sides. Each Cell Row is logically divided into four 16×8 *Cells*.

The four Cell Rows communicate with external resources via the open-source off-chip Dual Data Rate (DDR) full-duplex *RacEr_link_ddr* [8] located on the edges of the SoC, with

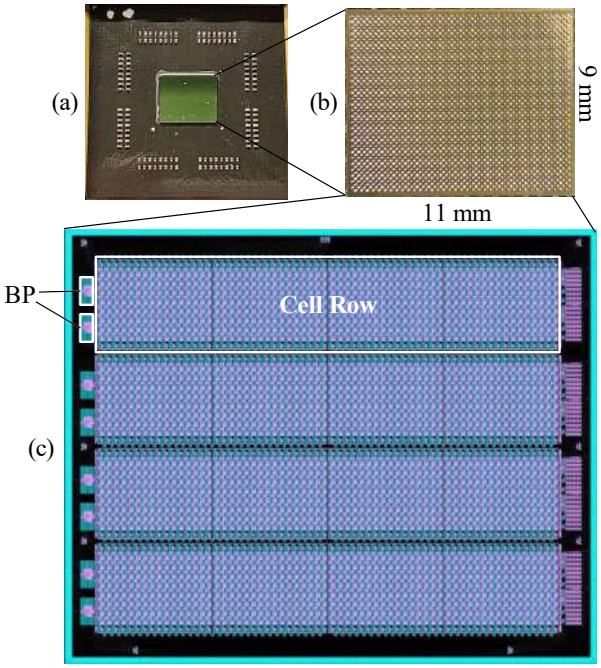
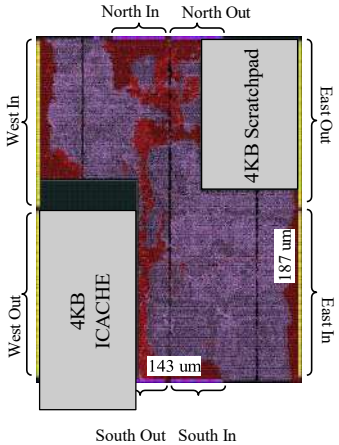


Fig. 1. (a) BGA 2116 flip-chip packaging, (b) 11×9 mm die photo, (c) chip floorplan. [7].

up to 1 *Gb/s*×512 *bits*÷8 = 64 GB/s bandwidth in total. The off-chip links have access to the memory resources, such as High Bandwidth Memory (UltrasoniK)2 and DDR4, via the off-chip Field-Programmable Gate Array (FPGA) bridge.

Figure 2 shows the compute tile layout and its area breakdown. Leveraging NoC symbiosis [2], the router (red) and core logic (pink) cells are co-placed within the tile. Ruche Networks on this chip demonstrate the high wiring density that can be achieved. Across the 187 um span of a tile, there are 2924 routing tracks in the two mid-level horizontal metal layers (these are above hardened Static Random-Access Memory (SRAM) macros), of which roughly 2416 tracks (~82%) are not blocked by power grids. Left and right sides of the tile each contain 1148 pins, occupying 47.5% of the available unblocked tracks. Despite the high routing density, Ruche Networks achieve a high area utilization of 85.3%, with physical-only cells (e.g. tap, filler cells) occupying the rest.



	Area (um²)	Area (%)
ICACHE	5785	21.6 %
Scratchpad	4359	16.3 %
Router FWD	**2245**	**8.4 %**
> Crossbar	452	1.7 %
> FIFO	1541	5.8 %
Router REV	**1372**	**5.1 %**
> Crossbar	230	0.9 %
> FIFO	891	3.3 %
Register Files	2789	10.4 %
ALU	356	1.3 %
FPU	2484	9.3 %
Other Logic	3453	12.9 %
Physical-only	3921	14.7 %
TOTAL	**26764**	**100.0 %**

Fig. 2. UltrasoniK compute tile floorplan and its area breakdown. Router cells (red), the core logic (pink).

IV SoC Integration Methodology

Implementing 2048 RISC-V cores on a monolithic die is challenging in terms of clock distribution and timing closure. Figure 3 shows three common methodologies for SoC inter-block clock distribution. System-synchronous designs (Figure 3(a)) such as [9], [10] require large, low skew clock trees that scale poorly with gate count and suffer from worsening variation. Standard globally-asynchronous locally-synchronous (GALS) designs (Figure 3(b)) employ a transmitter clock tree that extends into the receiver Clock Domain Crossing (CDC) block, requiring clock tree synthesis on the top-level.

We integrate the 99 mm² Ruche Networks SoC with a flexible, simple, verifiable and fast hierarchical design methodology—On-Chip Source-Synchronous Interconnect (OCSSI) (Figure 3(c))—which forwards clock together with data from transmitter to receiver and eliminates the top-level clock tree, so that top-level design is no longer dependent on block-level details. We demonstrate the methodology works in silicon. The end-to-end SoC integration scripts [11] are open-source for future reuse by the community.

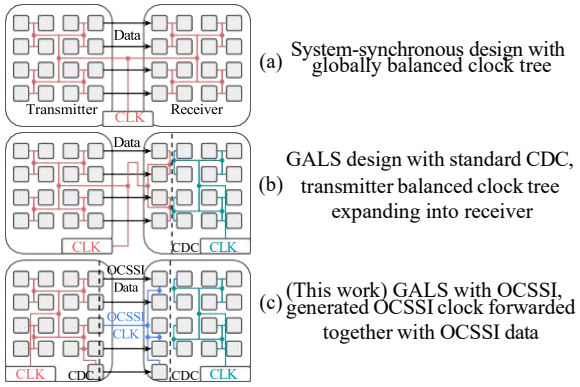


Fig. 3. Different SoC clock distribution methodologies. GALS with OCSSI makes large-scale SoC integration fast, flexible and verifiable with minimized dependencies between Intellectual Property (IP) blocks.

The SoC consists of tens of GALS Intellectual Property (IP) blocks. Each IP block has a standardized, width-configurable OCSSI interface that isolates the internal implementation details from the outside, as shown in figure 3(c). SoC designers have the flexibility to freely move an IP block without worrying about interconnect lengths, and can easily reuse or update an IP block while keeping the top-level timing closure clean. OCSSI transceiver design *RacEr_link_sdr* is at [8].

OCSSI simplifies top-level block implementation to drawing parallel wires. Figure 4 shows the SoC integration details. All the top-level links are source-synchronous; no clock tree synthesis is required. Four main IP blocks (each with 64×8 UltrasoniK cores) and 16 accelerator IP blocks are stitched together with 254 292-bit Short-OCSSIs that feature short lengths and tight pitch widths. Short-OCSSIs have massive bandwidth and 3-4 cycles of latency. Nine Input/Output (IO) IP blocks are connected to the main IP blocks via two 146-bit and 16 128- bit Long-OCSSIs, with lengths of 0.3 to 5.6 mm. Repeater nodes are evenly distributed on Long-OCSSI global routes to improve latency and signal integrity [12]. Long-OCSSI global

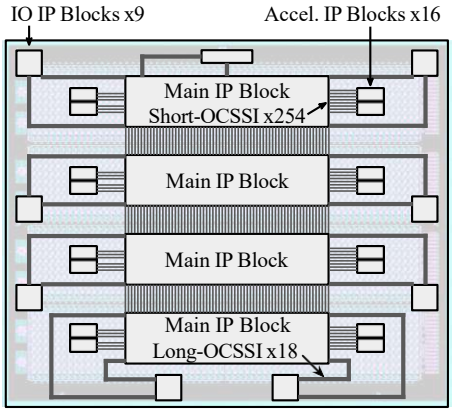


Fig. 4. SoC integration layout with 29 IP blocks and 272 OCSSIs.

routes are delay-matched to ensure clock-to-data alignments, where the delay value is mostly determined by the number of global buffers inserted on the path.

OCSSI shines in its outstanding verifiability. We sign off all 272 OCSS-Interconnects of 20 different lengths using Static Timing Analysis (STA), after integrating the routed top-level with all routed IP blocks. SPICE runs are not required in the top-level verification flow, and full-chip STA takes less than 10 hours per corner (multiple corners can be verified in parallel). In STA we create a generated clock on the 180-degree shifted clock signal, so that all OCSSI data paths become regular timing paths from the transmitter clock domain to the generated clock domain, which are synchronous to each other. In STA the OCSSI global routes (interconnect only, excluding transceiver logic) achieve frequencies higher than 2 GHz in all corners. The advantage of STA is that it takes Process, Voltage, and Temperature (PVT) variations into account, so that OCSS-Interconnects are guaranteed to work in all conditions. The OCSSI STA timing constraints are open-source [8], [11].

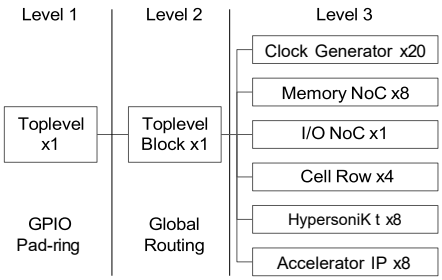


Fig. 5. SoC hierarchical design partitioning. Level 1 contains the physical pad-ring of the chip. Level 2 implements the global OCSSI routing among the leaf blocks. Level 3 has 49 leaf blocks spreading over the SoC. Levels are independent of each other and can be implemented in any order.

We leverage hierarchical design flow to reduce EDA runtime and iteration cost. Figure 5 shows the detailed partitioning of the SoC design. With OCSSI, parent blocks can be implemented without detailed information of child blocks, reducing the level-2 EDA runtime from days to hours. Moreover, the blocks can be implemented out-of-order without harming STA sign-off, parent blocks can be implemented before child blocks

TABLE I
BENCHMARK INPUT DATASET.

Benchmark	Input Dataset
AES	1 KB messages (2,097,152 copies)
SGEMM	512×512×512 (128 copies)
FFT	256×256 points (1024 copies)
Jacobi	512×512×512 grid
SpGEMM	Uniform random graph (64K nodes, 1M edges, 4 iterations)
BarnesHut	65536 bodies

are available (with a "dummy" placeholder block that only has OCSSI interface), and child blocks can be hot-swapped after parent blocks are implemented. This means all blocks in all levels can be implemented in parallel, reducing the iteration cost from end-to-end runtime to single-block runtime.

V RESULTS

A. Benchmark Performance Scaling on Silicon

We evaluated performance scaling across a diverse set of workloads spanning a wide range of compute and communication patterns. Table I summarizes the benchmarks and input datasets, while Figure 6 shows the speedup relative to single-core performance as core count increases from 1 to 2048. Scaling from 1 to 512 employs a single 64×8 core block mapped to two off-chip channels with UltrasoniKM2 attached. As off-chip bandwidth becomes saturated, performance scaling degrades, with SpGEMM exhibiting the highest memory intensity due to its irregular, pointer-chasing memory access patterns. In contrast, AES demonstrates near-ideal scaling due to its high compute-intensity. From 512 to 2048, scaling is achieved by expanding from one to two or four 64×8 core blocks, each with independent UltrasoniKM2 or DDR4 channels. As a result, performance scaling returns to a near-linear trend.

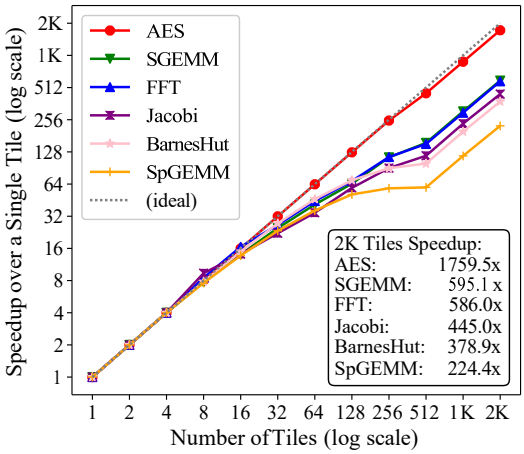


Fig. 6. Benchmark scaling from a single tile to 2048 tiles.

B. Ruche Packet Energy vs Distance

Figure 7 plots a trend in packet energy required to send a request and receive a response from remote tiles on a per-bit, per-length basis (pJ/bit/mm). The methodology used is similar to the one in [10]. A current consumption was measured as a subset of tiles executes an unrolled kernel loop that launches multiple remote requests without causing network

conflicts. Then, the difference in current measurements was taken between the remote access to itself and to the remote tiles. Measurements were taken with the activity factors of 0.25 (i.e. half of bits switching every cycle), approximately. Using the current differences and the rate at which packets were injected into the network, we derived the energy per packet, as the distance varies. Figure 7 shows its characteristic trend, where the Ruche packet energy dips every 3 hops. 2-D mesh energy is extrapolated from the first three data points. The dips are explained by the fact that using the long-range Ruche links to skip over 3 hops in one cycle costs less energy than hopping through individual routers using local links. On average, Ruche packet energy costs 0.054 pJ/bit/mm, 1.85× more efficient than 2-D mesh. According to the energy model developed in [13], the dynamic energy dissipated by NoC routers and long-range interconnects accounts for an average of 33.3% and 2.6% of the total, respectively.

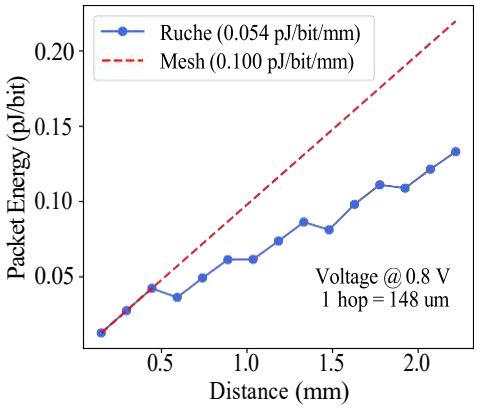


Fig. 7. Chip measurement: energy per packet, frequency at 0.5 GHz.

C. OCSSI Simulation Consistency and Performance

Figure 8 shows that Long-OCSSI works at up to 1.93 GHz on silicon at 0.8 V, which is remarkable for single-ended, voltage-mode-driven interconnects with fully standard cells, done with automated place and route, with up to 5.6 mm in length. Based on the wire-only STA results, it is safe to predict that if it had a faster OCSSI transceiver, it would achieve up to 3.04 GHz at 0.8 V on silicon. Figure 8 also shows that at 0.8 V, the measured maximum frequencies are close to the STA results of the OCSSI transceiver blocks, signaling good consistency between the STA simulation and the silicon. This demonstrates that our interconnect verification methodology and timing constraint scripts are effective for the OCSSIs.

OCSSI is not energy-thirsty, its power is reasonable despite the facts that it is long, using standard cells, and automat-ically placed and routed. It achieves 456 fJ/b with 5.6 mm length (0.58× compared to [14], lower is better, Long-OCSSI measured with simulation + silicon hybrid method at 0.8 V, 1.46GHz, α=0.25 (i.e. half of bits switching every cycle)).

VI Comparisons to Prior Works

A. NoC Comparison

Table II makes comparison with prior manycore designs. UltrasoniK is capable of operating at much higher fre-quency (1780 MHz at 1.0 V), despite having the long-range

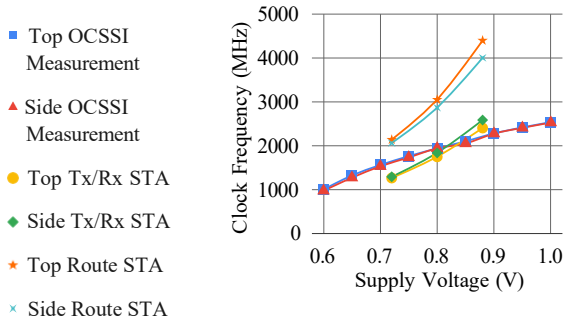


Fig. 8. Long-OCSSI performance consistency between STA and silicon. OCSSI measurements are conducted on silicon, from 0.6 V to 1.0 V; Tx/Rx stands for OCSSI transceiver STA; Route stands for OCSSI wire-only STA. STA results are simulated in slow, typical and fast corners.

Ruche links. NoC aggregate and bisection bandwidths are much higher in comparison, although this is helped by the fact that UltrasoniK has a larger die area and higher number of network routers. Routing density between UltrasoniK tiles is shown to be much higher compared to other designs in the same technology (e.g. DECADES, FlooNoC). ET-SoC-1 with concentrated mesh has significantly lower routing density compared to UltrasoniK despite using 1024-bit-wide channels and a 7 nm process. In addition, Ruche Networks reduces the packet latency, lower than the latency of Celerity, which was the previous fastest (x+y).

B. CoreMark Comparison

Table III compares the CoreMark score (CM). UltrasoniK attains CM of 8.7M, a 1.43× improvement over the current record holder, AMD's EPYC 9755 [22] at 6.0M – despite the fact that EPYC 9755 used a 4 nm process and 11.4× area. UltrasoniK outperforms Celerity [9], the previous RISC-V record-holder, in both compute density and energy efficiency, with improvements of 1.65× in CM/mm² and 1.15× in CM/W.

VII Related Work

FlooNoC [17] targets the high-bandwidth bulk data transfers required by Machine Learning (ML) accelerators by employ-ing very wide (~512-bit) links. Like Ruche Networks, it aims to better utilize on-chip wiring resources and adopts a low-complexity 2-D mesh router with minimal buffering and shallow pipelines. FlooNoC uses separate physical channels for the three AXI4 message classes (req, resp, wide) rather than virtual channels. However, wide meshes with multiple networks are not a cost-effective solution, as router area scales linearly with link width. In addition, core datapaths must be widened to match the NoC channels -— incurring further area overhead – otherwise they suffer de/serialization latency.

MemPool [23] provides a low-latency interconnect between processing elements and distributed L1 memory banks using hierarchical crossbars. It scales to 256 RISC-V cores and 1024 L1 banks organized in a three-level hierarchy (Groups, Tiles, Cores), enabling cores to reach any bank within five cycles in the absence of congestion and simplifying data placement for programmers. However, crossbar and wire routing presents a major scalability challenge, requiring large reserved routing corridors between clusters; resulting in either routing conges-tion or poor silicon area utilization.

TABLE II
NOC COMPARISON TABLE

Metric	KiloCore [15]	Piton [10]	Celerity [9]	DECADES [16]	FlooNoC [17]	ET-SoC-1 [18]	This Work[a]
Process	32 nm	32 nm	16 nm	12 nm	12 nm	7 nm	12 nm
Implementation	Silicon	Silicon	Silicon	Silicon	Post-Layout	Silicon	Silicon
Voltage (V)	1.1	1.0	0.60 - 0.98	1.2	0.8	0.4	0.8-1.0
Frequency (MHz)	1780	500	1400	911	1260	1000	**1490-1780**
NoC Topology	3×Mesh	3×Mesh	2×Mesh	3×Mesh	3×Mesh	1×CMesh	**2×Half Ruche**
Network Size	32×31	5×5	16×31	12×9	4×8	6×6	**64×32**
Aggregate BW (Tb/s)	388.4	11.3	361	68.2	104	122.9	**2572.6**
Bisection BW (Tb/s)	4.23	0.96	4.00	4.20	16.8	12.3	**53.2**
Routing Density[b] (bit/mm)	408	354	1002	712	1477	554	**4289**
Packet Latency[c] (cycle)	2*(x+y)+1	x+y+t+1	x+y	N/A	2*(x+y)	N/A	**(x+y)-2*[(x-1)/3]**
NoC Area (%)	9.0	2.9	7.8	N/A	6.9	N/A	13.5

[a] All measured at 0.8 V except for the max freq. [b] Includes both horizontal and vertical sides. [c] Latency for x horizontal, y vertical hops, and t turns.

TABLE III
COREMARK SCORE (CM) COMPARISON

Metric	Decades [16]	Cifer [19]	Celerity [9]	Epyc 9755	This Work
ISA	RISC-V	RISC-V	RISC-V	x86	RISC-V
Process	12 nm	12 nm	16 nm	4 nm	12 nm
Area(mm^2)	60.7[d]	11.64[d]	15.25	1129.6[a]	99
# Cores	60	22	496	128	2048
Freq(MHz)	911	1195	1400	2700	1632
Power(W)	N/A	1.792[c]	7.47	500[b]	69.25[e]
CM	153,959	27,116	812,350	6,065,430	**8,704,557**
CM/mm^2	2,536	2,330	53,269	5,370	**87,925**
CM/W	N/A	15,132	108,748	12,131	**125,698**

[a] CPU die area only [20]. [b] Power based on reported TDP [21]. [c] Nominal voltage active power. [d] eFPGA area removed. [e] Voltage at 1.0 V.

VIII CONCLUSION

The paper presents the UltrasoniK 2048-core RISC-V manycore SoC, which includes a 12 nm implementation of Ruche Networks, a novel, physically scalable, and tileable NoC topology. It introduces a flexible, simple, verifiable and fast SoC design methodology, OCSSI, which is essential for implementing a chip at this scale. The silicon shows strong performance scaling across a spectrum of real-world parallel workloads, from a single tile to 2048 tiles, and achieves the highest CoreMark score of any silicon to date.

REFERENCES

[1] D. C. Jung *et al.*, "Scalable, Programmable and Dense: The HammerBlade Open-Source RISC-V Manycore," in *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, 2024, pp. 770–784.
[2] VividSparks.tech, Internal Report, "Noc symbiosis ," *2020.*
[3] D. C. Jung, S. Davidson, C. Zhao, D. Richmond, and M. B. Taylor, "Ruche networks: Wire-maximal, no-fuss nocs," in *International Symposium on Networks-on-Chip (NOCS)*, 2020, pp. 1–8.
[4] Y. Ou, S. Agwa, and C. Batten, "Implementing low-diameter on-chip networks for manycore processors using a tiled physical design methodology," in *International Symposium on Networks-on-Chip (NOCS)*, 2020, pp. 1–8.
[5] J. Kim, J. Balfour, and W. Dally, "Flattened butterfly topology for on-chip networks," in *40th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 2007)*, 2007, pp. 172–182.
[6] B. Grot, J. Hestness, S. W. Keckler, and O. Mutlu, "Express cube topologies for on-chip interconnects," in *International Symposium on High Performance Computer Architecture (HPCA)*, 2009, pp. 163–174.
[7] VividSparks.tech, Internal Report, "HypersoniK: RISC-V Multicore Accelerator SoCs" 2021.
[8] VividSparks.tech, Internal Report ,"VividSparks STL libraries", 2020
[9] A. Rovinski *et al.*, "A 1.4 ghz 695 giga risc-v inst/s 496-core manycore processor with mesh on-chip network and an all-digital synthesized pll in 16nm cmos," in *2019 Symposium on VLSI Circuits*, 2019, pp. C30–C31.
[10] M. McKeown *et al.*, "Power and Energy Characterization of an Open Source 25-Core Manycore Processor," in *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2018, pp. 762–775.
[11] VividSparks.tech, Internal Report,:"RacEr manycore design", 2021.
[12] . Adler and E. Friedman, "Repeater design to reduce delay and power in resistive interconnect," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 45, no. 5, pp. 607–616, 1998.
[13] D. C. Jung and M. Taylor, "Evaluating Ruche Networks: Physically Scalable, Cost-Effective, Bandwidth-Flexible NoCs," in *Proceedings of the 52nd Annual International Symposium on Computer Architecture*, ser. ISCA '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 1035–1048.
[14] M. A. Anders *et al.*, "25.9 Reconfigurable Transient Current-Mode Global Interconnect Circuits in 10nm CMOS for High-Performance Processors with Wide Voltage-Frequency Operating Range," in *2020 IEEE International Solid-State Circuits Conference - (ISSCC)*, 2020, pp. 396–398.
[15] B. Bohnenstiehl *et al.*, "Kilocore: A 32-nm 1000-processor computational array," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 4, pp. 891–902, 2017.
[16] F. Gao *et al.*, "DECADES: A 67mm2, 1.46TOPS, 55 Giga Cache-Coherent 64-bit RISC-V Instructions per second, Heterogeneous Manycore SoC with 109 Tiles including Accelerators, Intelligent Storage, and eFPGA in 12nm FinFET," in *2023 IEEE Custom Integrated Circuits Conference (CICC)*, 2023, pp. 1–2.
[17] T. Fischer, M. Rogenmoser, T. Benz, F. K. Gu¨rkaynak, and L. Benini, "Floonoc: A 645-gb/s/link 0.15-pj/b/hop open-source noc with wide physical links and end-to-end axi4 parallel multistream support," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 33, no. 4, pp. 1094–1107, 2025.
[18] D. R. Ditzel and the Esperanto team, "Accelerating ML Recommendation With Over 1,000 RISC-V/Tensor Processors on Esperanto's ET-SoC-1 Chip," *IEEE Micro*, vol. 42, no. 3, pp. 31–38, 2022.
[19] T.-J. Chang *et al.*, "Cifer: A 12nm, 16mm2, 22-core soc with a 1541 lut6/mm2 1.92 mops/lut, fully synthesizable, cachecoherent, embedded fpga," in *2023 IEEE Custom Integrated Circuits Conference (CICC)*, 2023, pp. 1–2.
[20] "AMD Granite Ridge and Strix Point Zen 5 Die-sizes and Transistor Counts Confirmed," TechPowerUp, 2024. [Online]. Available: https://www.techpowerup.com/324562
[21] "AMD EPYC 9005 Series Processors Datasheet," Advanced Micro Devices, Inc., Santa Clara, CA, USA, 2025. [Online]. Available: https://www.amd.com/content/dam/amd/en/documents/epyc-business-docs/datasheets/amd-epyc-9005-series-processor-datasheet.pdf
[22] M. Larabel, "Coremark 1.0," accessed: Aug. 13, 2025. [Online]. Available: https://openbenchmarking.org/test/pts/coremark
[23] S. Riedel, M. Cavalcante, R. Andri, and L. Benini, "Mempool: A scalable manycore architecture with a low-latency shared l1 memory," *IEEE Transactions on Computers*, vol. 72, no. 12, pp. 3561–3575, 2023.