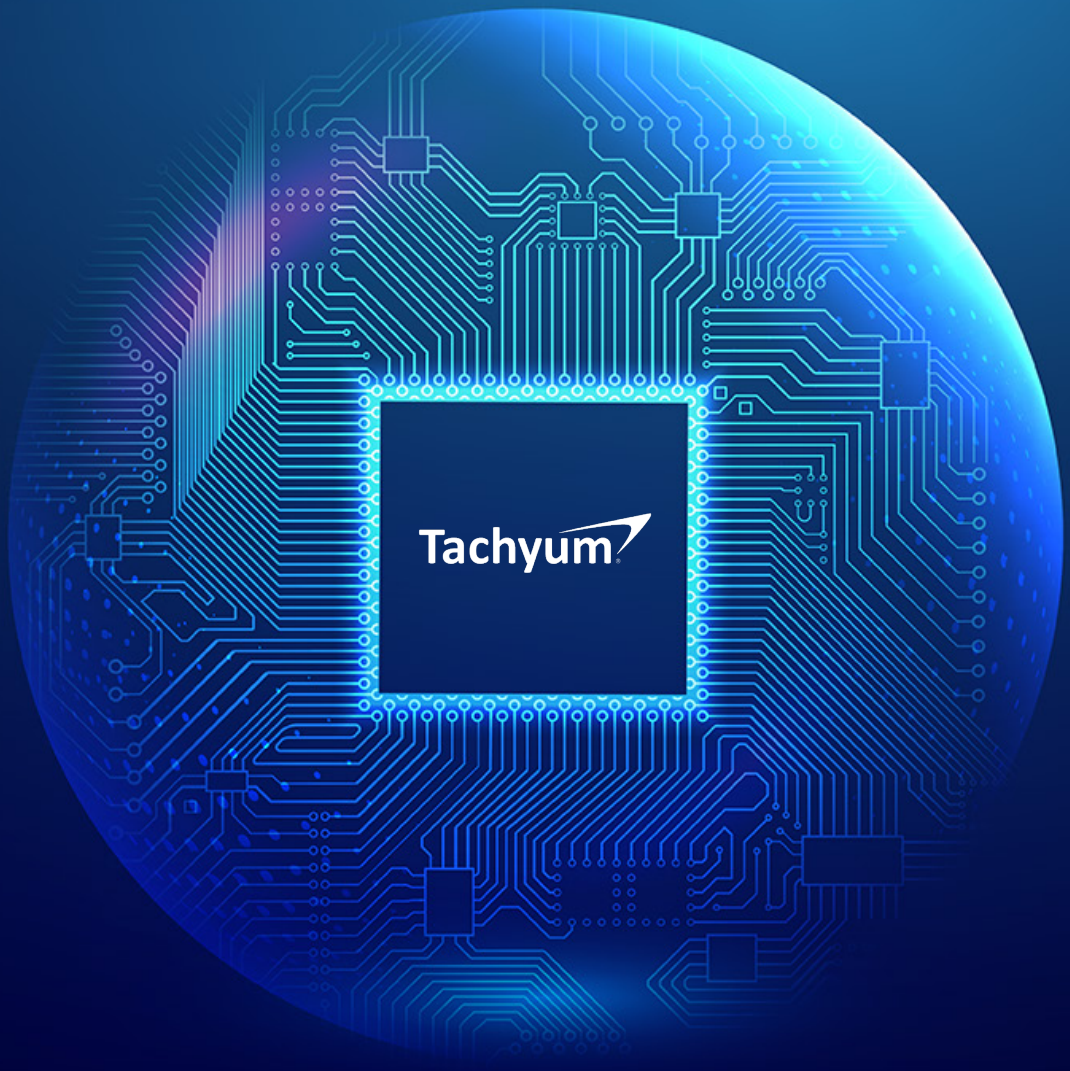
Tachyum Prodigy Feature Upgrades

November 2025



Introduction

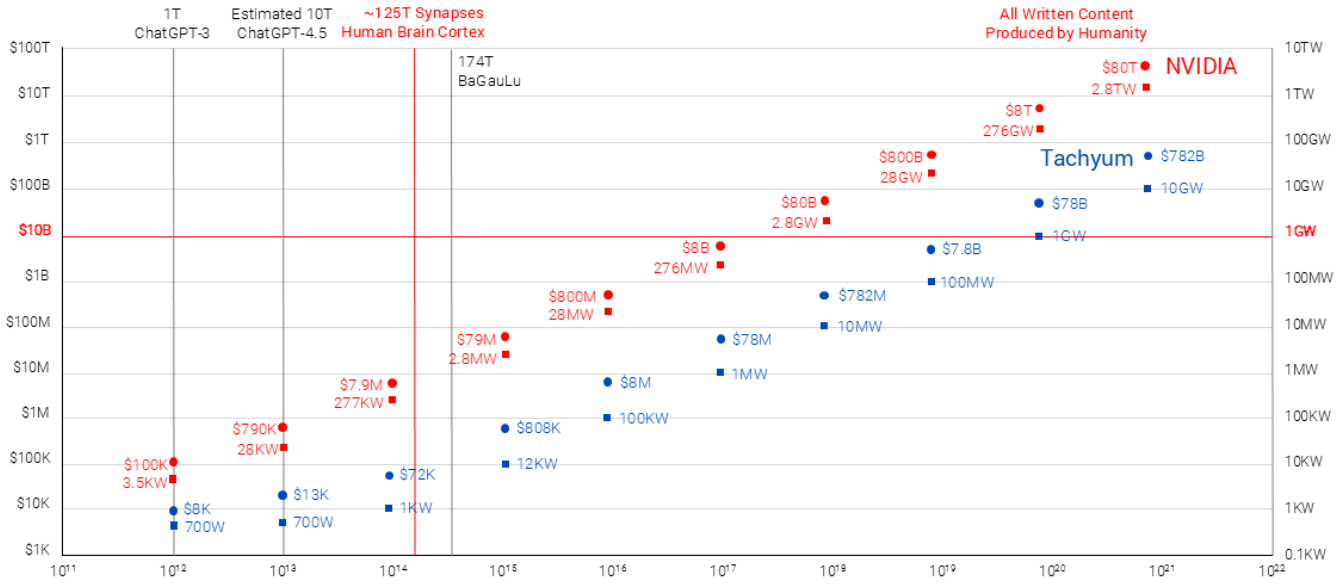
This document provides a technical summary of the Prodigy feature changes from Prodigy's previous 256-core product definition to the recent upgrade to 1024 cores, the performance increase for the 1024-core product at the device level as well as within the different subsystems, and an explanation of many of the engineering tradeoffs that our architects considered to finalize the optimal product definition.

The target audience includes analysts, lead customers, and key partners. In addition to the technical explanation of the feature enhancements and resulting performance improvements.

Due to the dynamic, ever-changing market landscape for both CPUs and GPGPUs. Tachyum's enhancements to the product definition ensure a highly differentiated, disruptive, cutting-edge solution for cloud, AI, and HPC workloads in the 2026/2027 timeframe. Overlayed on top of the technical solution is a first-rate team of experienced, world-class engineering talent that is dedicated to driving the market disruption that we will be delivering with Prodigy.

Our level of confidence is very high for design and market execution for the newly enhanced Prodigy. In addition to our top-notch engineering team and key partnerships, we now have $220M in Series C funding which is sufficient to enable development, fabrication, and production.

Next steps include first customer samples in Q1, 2027.



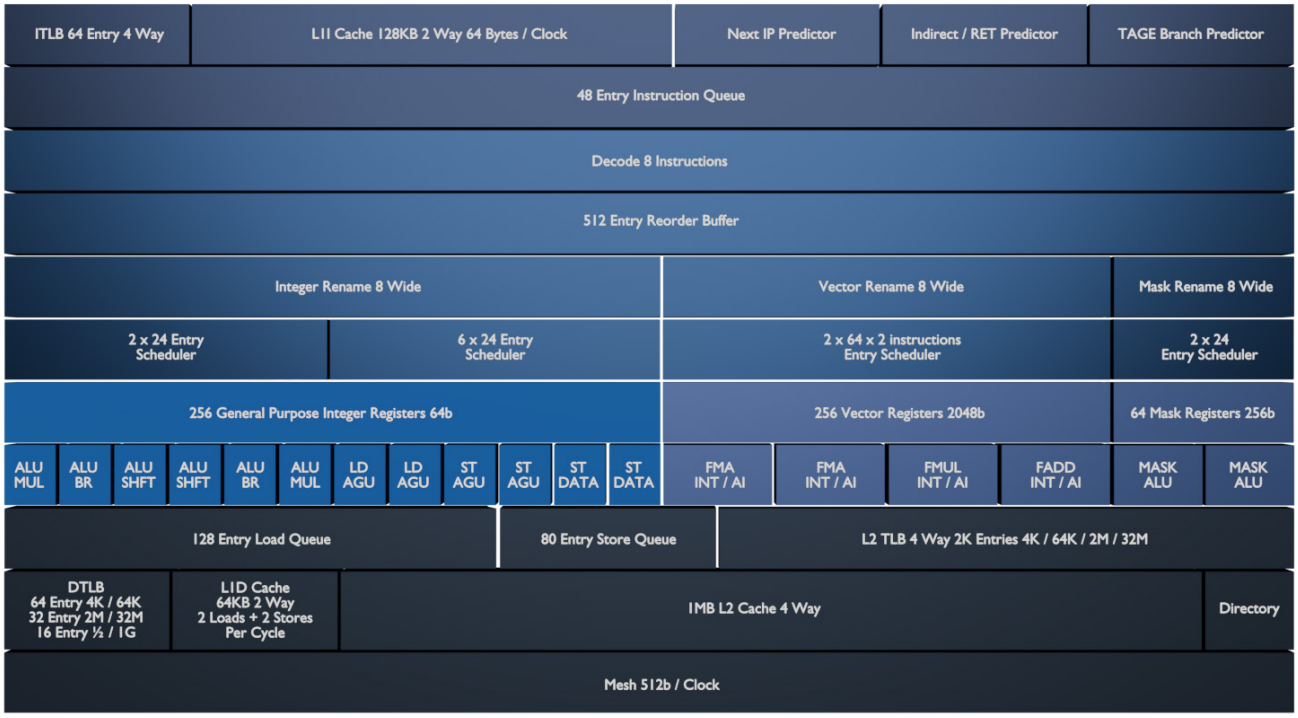
2nm Prodigy Improvements

The improved Prodigy in 2nm increases integer performance by 5x, AI performance by 16x, DRAM bandwidth by 4x, chip to chip bandwidth by 4x, and I/O bandwidth by 4x. Additionally, scalability increased 4x by expanding chip-to-chip hardware coherency from 4 sockets to 16 sockets. Power efficiency for the same performance was increased by 2x, while cost per core was reduced.

Core Microarchitecture

The 5nm Prodigy was an out of order execution machine and the 2nm version improved out of order execution and performance. The diagram below shows the 2nm Prodigy core architecture.

For training LLMs, the FP4 format carries several notable improvements:



16x AI Performance Increase

The AI performance of the chip increased by 16x, and the AI node performance increased by 32x, by increasing number of sockets supported by 2x from 4 to 8.

The number of cores increased by 4x from 256 to 1024. The performance per AI multiply-add units increased by 2x due to enlarging the matrix size from 1Kbit back to 2Kbits. The number of FP8 multiply-add matrix units increased by 1.5x to 3, and the number of FP4 and TAI (Tachyum AI) units increased by 2x to 4.

5x Integer Performance Increase

The integer performance increased by over 5x. The 4x increase comes from having 4x more cores and more than a 25% increase from the Instruction Per Clock (IPC).

IPC increases came from replacing branch prediction with a TAGE predictor, 4x the increase fetch width from 16 bytes to 64 bytes per clock, 3x increase in the size of the Reorder Buffer (ROB), improved dependent load after store executed too early, prefetcher improvements, data cache and other smaller improvements.

More than 5x HPC Performance Increase

Approximately 5-6x HPC performance increase comes from having 4x more cores and over 25% of the increase comes from having better scalar performance, selecting the optimal vector length, and a 4-issue vector unit with a shorter data path length.

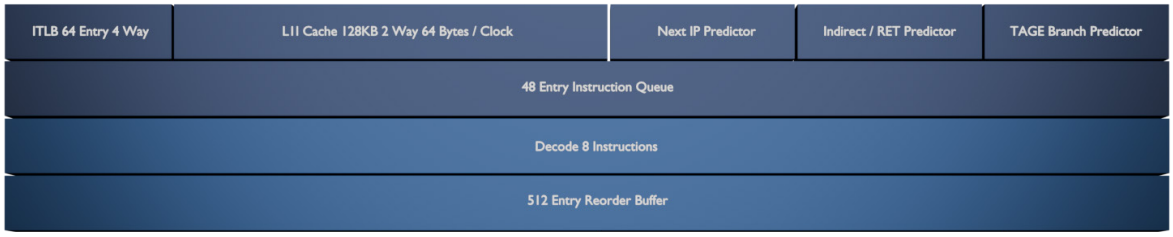
4x Memory Bandwidth

The memory bandwidth increased over 4x from 16 x DDR5 7.2 GT/s * 8 bytes = 820 GB/s to DDR5 MRDIMM 3.0 with17.6 GT/s * 24 DRAM controllers * 8 bytes * 2 bandwidth amplification is 6.76 TB/s. For AI, TAI2 now allows inference with bandwidth of equivalent to 13.5 TB/s.

Front End

The Front End (FE) job is to supply the core with instructions without being bottlenecked to 8 instructions per clock instruction decoder, rename, Reorder Buffer (ROB), and many execution units.

The FE should supply instructions at more than 2x the Instruction Per Clock (IPC) execution rate, because it must also fill the Reorder Buffer (ROB) to extract parallelism. The FE was widened by 4x from 16 bytes per clock to 64 bytes per clock with corresponding improvements in branch prediction.



Instruction Fetch

The instruction fetch was widened by 4x from 16 bytes per clock to 64 bytes per clock. Branch prediction switched to a TAGE branch predictor, a collapsing Instruction Queue (IQ) was added, and the penalty for Front End (FE) misprediction was significantly reduced.

Branch Prediction

Updating the branch prediction with a TAGE branch predictor reduced branch misprediction by >2x. The branch predictor can predict up to 8 branches per clock.

The number of potentially taken branches in the Next Instruction Pointer (NIP) increased accordingly from 1 per 16 bytes to 4 per 64 bytes.

With the increase in the Reorder Buffer (ROB) depth, the return stack predictor depth was also increased, as well as the return stack aliasing being reduced.

Instruction Queue

Because the fetch unit can fetch up to 16 instructions per clock, it requires the Instruction Queue (IQ) between the front end and the instruction decode. It takes 2 clocks to stop the Front End (FE) on the scheduler resources overflow, which sets the IQ size to 48 instructions.

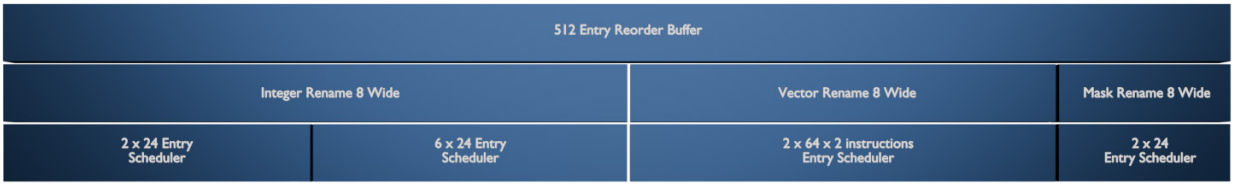
The instruction queue collapses instructions to avoid wasting rename and Reorder Buffer (ROB) bandwidth and improves utilization.

8-Wide Instruction Decode

The instruction decode is widened by 2x to 8 instructions per clock. The rename stage is 2x widened and renames 8 instructions per clock. The 8 instructions per clock can be placed into the Reorder Buffer (ROB). Retirement was widened by 2x to retire 8 instructions per clock.

Schedulers

The scheduler depth increased from 15 to 24, and together with efficiency improvements from data cache miss handling, load bank conflict scheduling, and other improvements, the effective depth of schedulers more than doubled.



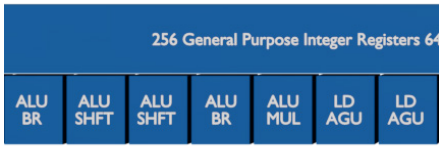
The credit-based stall on depleting resources like physical registers, Reorder Buffers, and other resources created problems for an 8-issue machine and was replaced with a stall on overflow, thus dramatically improving system efficiency.

Execution Units

The integer execution units increased from 4 to 6, and vector execution units from 2 to 4. The 2 loads, 2 stores, and 2 mask units remain unchanged.

Integer Execution Unit

The larger Reorder Buffer (ROB) extracts more parallelism, enabling the benefit of more execution units. The number of integer execution units increased from 4 to 6, and 2 of the 6 units support multiplication and certain special operations.



Mask Execution Unit

Masks have been extended to support micro-exponents for AI. The number of mask execution units remained at 2.

Vector Unit

The 2-issue vector unit data path was modified to a 4-issue unit, but with a 512-bit data path for integer, floating-point, and non-AI operations. The vector unit is the most power-hungry block in the design and doubling performance with a significant increase of power and thermal envelope did not make sense, as Prodigy already has high vector performance.

The vector length can be set to 256, 512, 1024, or 2048 bits. When the vector length is set to 1024 or 2048 bits, the operations are executed over multiple cycles. If the existing code base is optimized for 512-bit vectors, setting the vector length to 512b may be optimal. Vector operations can be mixed with AI operations only if the vector length is set to 2048 bits.

Floating-Point Units

The Floating-Point (FP) Multiply-Add (FMA) unit was modified to provide results of FP multiplication with shorter latency. The FP adder with reduced latency was added. The Square Root unit Radix-16 was added to the existing approximation unit. The divider has been upgraded from Radix-4 to Radix-16 to reduce latency.

The scalar FP performance improved when dual-issue Vector Units (VU) increased to quad issue. In addition, scalar FP performance is being improved by adding additional instances of scalar units to the quad-issue vector unit.

AI Unit

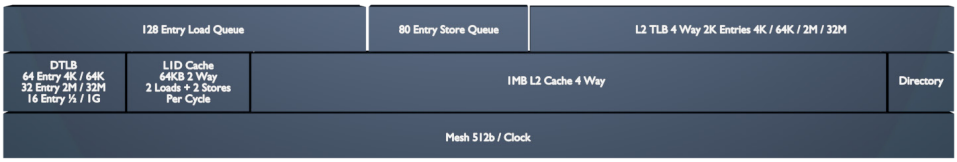
The AI unit has 4 matrix-multiply-add units with 2Kb matrices. They support both dense and sparse INT16, INT8, INT4, TF32, BF16, FP8, FP4, as well as Tachyum proprietary TAI formats. The 2Kb matrices doubled performance across the data types. The number of execution units supporting INT16, INT8, INT4, TF32, and BF16 did not increase. The matrix-multiply-add units for FP8 increased to 3, and for FP4 and TAI to 4.

The mask registers for AI operations are used as a micro-exponent. The number of architectural and mask registers has been increased to reflect that use. The architecture supports both OCP as well as Tachyum proprietary micro-exponent matrix formats.

Support for sparsity separate from data was added in addition to the existing Tachyum AI (TAI) data types. Separating sparsity to separate registers allows support for existing models as-is without re-quantizing them to TAI data formats, which are more efficient.

Memory Subsystem

The memory subsystem was improved considerably. The virtual and physical address spaces were both enlarged to add support for large CXL memory pools behind CXL switches. The 64KB page size mode was optimized for mTHP. The bandwidth and capacity of memory were greatly expanded.



64KB Page Support Is Default

Prodigy supports both 4KB and 64KB pages and supports default out-of-box Linux and FreeBSD support for 64KB pages. Tachyum will run regressions for 64KB page support in operating systems and applications.

4KB page size support for customers, if required, could be provided with an agreement between the customer and Tachyum.

Load/Store

The 2 loads and 2 stores per clock did not increase. Unaligned vector loads and stores can be executed in a single cycle if they do not cross a page boundary.

The load and store queues were increased and conflict load after a store executed early predictor was dramatically improved. Schedulers were modified to reduce the data cache miss penalty and bank conflict penalties, along with other smaller improvements.

Data Cache

The bank conflict reduction and data cache miss penalty reduction improved the IPC. With larger issue width and reorder buffer size, loads in program order before stores are executed earlier than these conflicting stores, causing considerable increases to pipeline flushes. Replacing with a much more powerful predictor reduced the penalty more than an order of magnitude and increased the IPC close to 10%

L2 Cache

The L2 cache prefetcher was improved to track more complex memory access miss patterns, improving performance and reducing bandwidth overhead.

Chip

The process shrink from 5nm to 2nm allows hitting the clock speed at a lower voltage, thus reducing power consumption versus the 5nm design. Reducing chiplet die size improves yield and cost despite more expensive 2nm wafers versus the 5 nm process. The power consumption reduction is very important as multiple chiplets will occupy one package.

From Reticle Sized Die to Chiplets

As the initial design on a 5 nm process grew over time from 128, through 192, to 256 cores, it reached reticle size. It was split into 2 halves, so if there is a defect in one half of the die, the second half can be used for lower performance SKUs. That capability was retained in the chiplets.

The 2 nm wafers are more expensive than the 5nm wafers. Splitting a reticle-size die into a smaller one, let's call them chiplets, improves yield and reduces cost to offset the higher cost per wafer. The 256-core chiplet design provides optimal benefits between cost and flexibility of product SKUs.

On Chip Core to Core Mesh Interconnect

With 4x more cores, 4x increase in memory bandwidth, and 4x increase in I/O bandwidth, core-to-core mesh interconnect bandwidth is increased by 4x. It doubles the speed of the mesh by pipelining the mesh to run at the core clock, up from half of the core clock, retaining the same absolute delays. Now, as before, one hop takes core cycles, since the mesh is now running at the core clock instead of half of the core clock.

To get 4x bandwidth increase, the width of the mesh was increased from 32 bytes to 64 bytes per clock. So, the mesh is now 2x faster and 2x wider.

DRAM Subsystem

It's necessary to ensure that cores are not starved for memory bandwidth. Prodigy supports MRDIMMs to more than double the bandwidth from 6.4 GT/s to up to 17.6 GT/s, in addition to increasing the number of channels from 16 to 24, thus increasing memory bandwidth by 4x to feed 4x more cores.

The direct-attached DRAM memory capacity increased 8x from 4 sockets * 16 channels * 256 GB DDR5 TSV2 RDIMM to 16 sockets * 16 channels * 512 GB DDR5 DDP TSV2 MRDIMM. Memory capacity could be expanded using CXL memory to much higher capacities.

Chip to Chip Interconnect

With 4x increase of memory bandwidth and 4x increase of I/O bandwidth, the Chip-to-Chip interconnect bandwidth had to be scaled by 4x too. This is achieved by upgrading SERDES from 112G to 224G and doubling the number of SERDES allocated for Chip-to-Chip interconnect.

Doubling the number of SERDES together with an extended directory allows an increasing number of sockets supported from 4 up to 16, thus increasing the system performance by 16x. Tachyum leverages UALink 200G IP.

Directory

The directory was extended from supporting 1,024 cores to a much larger number and includes support for CXL 3.2 and future extensions. For CXL sharers, all cache lines in DRAM can be accurately tracked.

I/O Subsystem

With a 4x increase in cores, the PCIE I/O subsystem bandwidth is also increased by 4x by replacing PCIE 5.0 32Gb/s to PCIE 7.0 128Gb/s per lane.

The PCIE IP integrates CXL 3.2 protocol to dramatically increase memory capacity per system by orders of magnitude.

Chip Packages

There are distinct markets addressed by Prodigy. One is the server market, and another is the AI/HPC market.

The CPU market is dominated by air-cooled systems and high-volume, with many industry standard form factors fitting different system and application requirements. To ensure easy and fast adoption, Prodigy focuses on accepted dimensions, power, thermal specifications, scalability, features, and cost segmentation. This market is served by socketed systems with Prodigy's 120 mm x 63 mm Standard package having 6,944 balls. This package is optimized for air-cooled designs but also allows liquid cooling solutions.

Another market is AI/HPC, which is moving from Air-Cooled to Liquid-Cooled systems. Chips in this market are larger, have higher performance, higher power, higher TDP , higher memory bandwidth, and are more expensive. This market is served by socketed systems with Prodigy's 120 mm x 150 mm High-Performance package having 23,009 balls. This package is optimized for liquid-cooled designs but allows air-cooled solutions at lower performance.

Product SKUs

A single Prodigy tape-out allows customers to create a wide range of product Stock Keeping Units (SKUs) covering a wide range of performance and applications. The Prodigy SKUs are summarized in the table below, and specify the SKU features, including whether the SKU is built with the 120mm x 63mm Standard package or 120mm x 150mm High-Performance package.

Product SKU	Number of Cores	Max Frequency (GHz)	Scalability	DDR5 Controllers	DDR5 Speed	PCIe 7.0 Lanes	TDP	Target Markets / Applications
T241024	1024	6.0	8S	24	17600	128	1600	Top-End HPC, Big AI
T24768	768	5.0	8S	24	17600	128	1000	Entry HPC, Big AI, Crypto, Digital Currency
T16512	512	6.0	16S	16	12800	128	800	Analytics, Big Data, Crypto, Digital Currency
T16448	448	5.5	16S	16	12800	128	645	Analytics, Big Data, In-Memory Databases
T16384	384	5.5	2S	16	12800	128	550	Cloud, Analytics, Big Data
T16320	320	5.0	2S	16	10700	128	420	Cloud, Edge/Telco
T16256	256	4.5	2S	16	10700	128	300	Cloud, Databases, Edge/Telco
T8256	256	4.5	1S	8	10700	96	300	Cloud, Databases, Edge/Telco
T8128	128	4.5	1S	8	9600	96	150	Cloud, Databases, Storage
T896	96	5.5	1S	8	9600	96	140	Cloud, Databases, Storage
T464	64	4.0	1S	4	6400	48	70	Entry Cloud, Storage
T432	32	3.5	1S	4	6400	24	30	Low Power, Hosting

Single-Socket Standard Package

The single-socket SKUs in Standard packages are the lowest cost chips. They enable workstations, high density web servers, and high-density cloud applications. These will typically be air-cooled.

They come with 8 DRAM channels with 1 or 2 DIMMs per channel and 256, 128 and 64 core options. The entry level lowest power and lowest cost SKUs come with 4 DRAM channels with 1 or 2 DIMMs per channel and 128, 96, 64 and 32 cores.

Multi-Socket Standard Package

The multi-socket in the Standard package are the most flexible and likely the highest volume SKUs. Most systems will be targeting 19" standard racks. While most volume likely will be air cooled, Tachyum will qualify liquid-cooled solutions from partners.

They come with 16 DRAM channels with 1 or 2 DIMMs per channel, and 512, 256 or 128 cores. They support 1, 2, 4, 8 and 16 socket systems. Performance is adjusted based on TDP and the type of cooling.

Multi-Socket High-Performance Package

The multi-socket High-Performance package is built for SKUs with 1024 or 768 cores and 24 DRAM channels supporting 1 DIMM per channel. This will focus on 4 and 8 socket liquid cooled systems in wide racks, but an air-cooled option with 4 sockets configured for 19" racks would also be possible.

The primary expected applications are High Performance Computing (HPC) and Artificial Intelligence (AI) solutions.

Software

Tachyum provides not only out-of-the-box native system software, operating systems, compilers, libraries, debuggers, and AI infrastructure and frameworks, but also many applications. In addition, it allows running unmodified Intel/AMD x86 binaries and mixing them with native applications. This ensures that Tachyum systems can be deployed by customers from day one.

Operating Systems, Virtualization and System Software

Tachyum provides its partners with source code firmware and software to build systems. Partners can customize code. System software includes, for example System Management Mode (SMM), UEFI, OpenBMC, GRUB, systemd-boot, Error Detection and Correction (EDAC) and others.

Operating systems include native Linux, SELinux, and FreeBSD. Tachyum provides native virtualization and associated software, including Linux KVM, Xen, Docker, Kubernetes, Samba, Ansible, and others.

Intel/AMD x86-64 Binaries

Tachyum allows running unmodified Linux application binaries through dynamic binary translation. Linux is modified to transparently run these binaries, and application packages can mix native binaries with x86 binaries. The x86 binary can be launched directly, and Linux modified by Tachyum will recognize a foreign binary and transparently launch dynamic binary translation. For example, it was demonstrated to run a proprietary x86 database with a native Apache web server without needing to modify scripts.

Tachyum also demonstrated dynamic binary translation for ARM and RISC-V binaries. Customers who require running unmodified ARM and RISC-V binaries can contact Tachyum .

Compilers, Languages and Debuggers

Tachyum provides the most recent GCC and LLVM compilers. Supported languages include C/C++, Clang, Erlang, Fortran, Go, Java JVM, Lua, OpenJDK, Perl, PHP, Python, R language, Ruby, TCL, Rust and others.

Libraries include Glibc, gRPC, OpenMP, Open MPI, Eigen Library, LAPACK, Real and Complex Eigensolvers for Large linear systems, FFT Library, ODE/OPE numerical solvers and others. Debuggers include gdbsim, GDB, OpenOCD, KGDB, JTAG Debugger, and others. Optimization tools include AutoFDO, PERF Instruction Profiling, PERF Counters, and others.

Artificial Intelligence (AI)

Tachyum supports native PyTorch, TensorFlow, Stochastic Rounding for BLAS-GEMM, Mixed precision training, Compression algorithms, Computer Vision, Object detection and semantic segmentation models, NLP Transformer Models, and others.

Applications

High Performance Computing (HPC) support includes DeepMD, Quantum Espresso, LAMMPS, SLURM, LINPACK Benchmark, HPL (High Performance LINPACK), and others. Native applications include Apache Hadoop, Apache Spark, SVN, Apache Web and Email, RabbitMQ, Postfix, Dovecot, Ceph, and others.

Native databases include MariaDB, PostgreSQL, MongoDB, SQLite, RocksDB, and others.

Early Adopters Support

For early adopter software development, Tachyum has servers hosting QEMU emulation of Tachyum systems on standard Intel/AMD servers. Performance is comparable to a 2.2 GHz server processor and allows customers to work in native environments instead of in cross platform development environments.

Customers who try to performance-tune applications there are very fast C-models running equivalent to over 200 MHz frequency. If customers want to understand the intricate details of out-of-order execution, they could use a transactional accurate C-model.

Finally, the most precious and most expensive tool is Tachyum's hardware FPGA emulation system running at 50MHz, on which Tachyum engineers can run customer applications.

Support in Porting Customer Applications

Tachyum has a network of partners to help customers port their proprietary application to Prodigy platform.

Reference Motherboards

Tachyum, with its partners, will provide motherboards to cover a wide range of customer cases.

1-Socket

Single socket motherboards for desktops, workstations, and certain servers can accept Standard packages with 8 DRAM channels with 2 DIMMs per channel, 256 or 128 cores, or 4 DRAM channels with 2 DIMMs per channel for 128 or 64 or 32 cores. It is an air-cooled design with heatsinks.

2 Sockets

The 2-socket motherboards are expected to use narrow packages with 2 sockets behind each other and each socket connected by 16 DRAM channels to 1 or 2 DIMMs per channel. Most common systems are expected to be air-cooled in 1U to 3U chassis, depending on TDP limit, but liquid-cooled would be possible for a 1U chassis.

4 Sockets

The 4-socket motherboard uses Standard packages with 2 sockets side by side and 2 other sockets behind them to fit into standard 19” rack. Each socket is connected to 16 DRAM channels with 1 DIMM per channel. Most common systems will be air-cooled in 1U to 3U chassis, depending on TDP limit, but liquid-cooled would be possible for 1U chassis.

8 Sockets

The 8-socket motherboards will be provided by Tachyum partners using Standard packages to fit into standard 19” rack with air cooling. They would be built as 4 dual sockets motherboards. Liquid cooling would be possible.

The 8-socket motherboards for wide racks using High-Performance packages would be composed of 2 motherboards in a 1U liquid-cooled system. Air cooling options, depending on the TDP limit, may also be possible.

16 Sockets

The 16-socket system targets wide racks composed of 4 motherboards, each with 4 sockets connected by cable for wide racks. Each Standard package supports 16 DRAM channels with 1 DIMM per channel. Most common systems are expected to be 1U liquid-cooled, but air cooling would be possible for 3U or larger servers.

Solutions

Tachyum Prodigy offers the flexibility of building various solutions, with the most common reference designs created with Tachyum and its partners. Solutions allow air or liquid cooling, and fit into standard 19” racks or wide racks, such for example Open Rack Wide Systems (ORW).

It offers scalability from 1 to 16 Standard sockets or 8 High-Performance sockets. Applications span from conventional servers to converged conventional + AI servers, to High Performance Computing (HPC), or AI systems.

Liquid Cooling

Liquid cooling is getting increasing traction from the current estimated 40% of High-Performance Computing (HPC) and Artificial Intelligence (AI) servers to over 60% in two years.

The 2 and 4-socket Standard package motherboards for standard 19" racks 1U systems will support third party liquid cooling solutions, which will be tested by Tachyum to provide the approved vendor list.

Tachyum will optimize liquid cooling solutions for 4 and 8-socket High-Performance packages for 1U systems to extract maximum performance.

Air Cooling

The air-cooled standard 19" rack systems will have heatsinks optimized for full performance 2 and 4-socket systems in 3U enclosures.

The air-cooled standard 19" rack systems will have heatsink-optimized for single sockets and dual-socket high density systems. The 1U air-cooled systems will have performance-limited and core-count limited by TDP.

The dual-socket High_Performance package motherboards may also be used with air-cooled systems using heatsinks for 3U enclosures.

Standard 19" Racks

The 2-socket and 4-socket High-Performance package air-cooled chips and motherboards using standard power supplies are primarily targeting standard air-cooled 19" racks. The heatsinks will be optimized for 1U and 3U chassis. The same motherboard would also support liquid cooling servers in 1U form factors.

Open Compute Project High-Density Dual-Socket Solutions

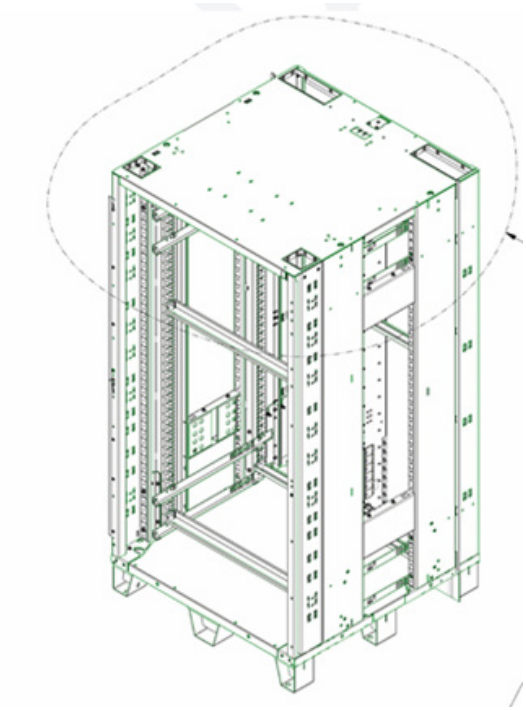
Tachyum partners will provide solutions for high-density dual-socket solutions. The Standard packages with 8 DRAM controllers with 64, 128 or 256 cores are expected to be used to serve this market.

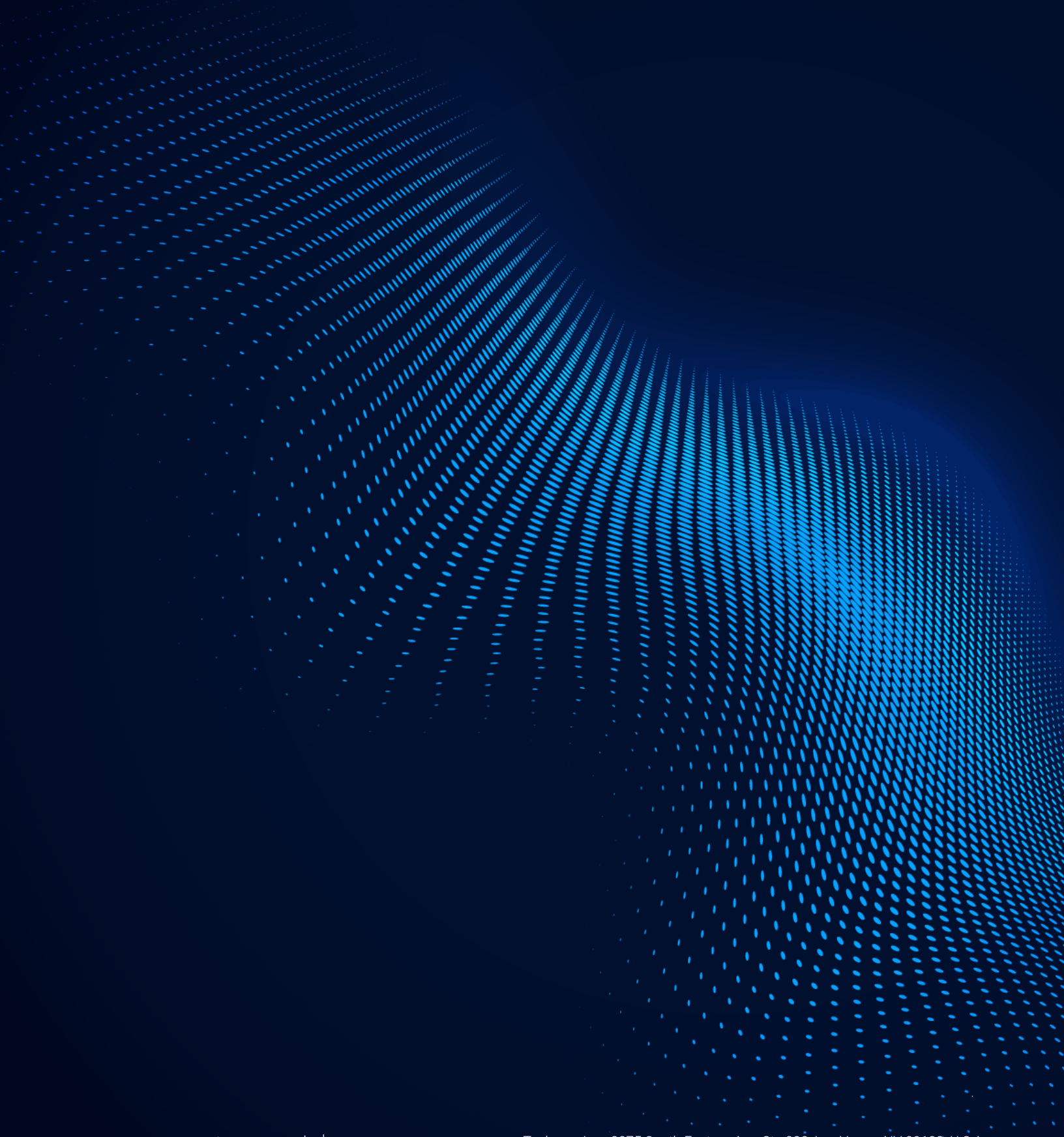
Open Compute Project High Density Single Sockets Solutions

Tachyum partners will provide solutions for high-density single sockets solutions. The narrow packages with 4 DRAM controllers with 32, 64 or 128 cores are expected to be used to serve this market.

Open Rack Wide Systems

The updated Prodigy was designed to take full advantage of the Open Rack Wide (ORW) Open Computer Project (OCP) standard. It mechanically allows building up to 16-CPU socket systems, up from today's 8-socket systems used in some supercomputers.





Tachyum

www.tachyum.com

Tachyum Inc., 8275 South Eastern Ave, Ste 233, Las Vegas, NV 89123, U.S.A.
Tachyum s.r.o., Karadžičova 14, CBC IV., 3rd floor, 821 08 Bratislava, Slovakia

Prodigy New Specs v1.0_251117