Trends and Infrastructure for AI

E. Jan Vardaman, President and Founder

- TRACK INNOVATION
- IDENTIFY TRENDS
- ANALYZE GROWTH
- INFLUENCE DECISIONS

RELEVANT, ACCURATE, TIMELY

techsearchinc.com

© 2025 TechSearch International, Inc.

TechSearch INTERNATIONAL

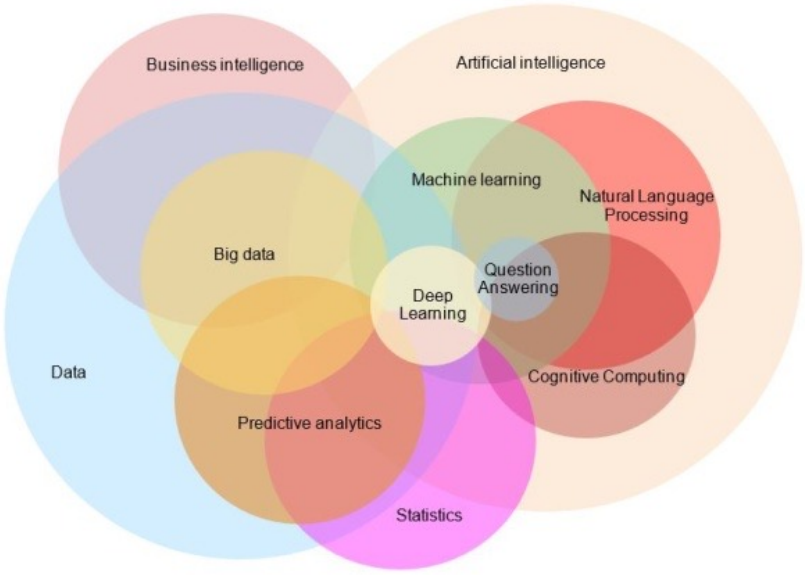
Outline

- **AI for data center**
  - Package examples
  - Assembly revenue projections for AI
  - Demand for CPO
- **Edge AI**
- **AI infrastructure growth**

© 2025 TechSearch International, Inc.

Artificial Intelligence

- **Artificial Intelligence (AI): Theory and development of computer systems able to perform tasks that normally require human intelligence**
  - Visual perception and pattern recognition, Speech recognition, Decision-making, Natural language processing and translation
- **Machine Learning: Branch of AI in which computers learn from data without human assistance**
- **Deep Learning: Type of machine learning that trains a computer to perform human-like tasks**
  - Recognizing speech, identifying images, or making predictions
  - Sets up the parameters about the data and trains the computer to learn on its own by recognizing patterns using many layers of processing



Source: IBM.

© 2025 TechSearch International, Inc.

AI for Data Center

techsearchinc.com

© 2025 TechSearch International, Inc.

TechSearch INTERNATIONAL

AI Package Examples

Company Product	Package Size (mm) (Substrate Construction)	Package
AMD Instinct MI200 (with chiplets)	78 x 70 (5-2-5)	Elevated Fanout Bridge (EFB) for GPUs + HBM, connected to laminate build-up substrate
AMD Instinct MI300 (with chiplets)	75.4 x 72 (8-2-8)	GPU +CPU logic-on-logic stack with SoIC + HBM on Si interposer with DTC
Intel Ponte Vecchio (with chiplets using Foveros and EMIB)	77.5 x 62.5 (11-2-11)	Foveros with logic + HBM, connected to embedded silicon bridge in laminate substrate
Nvidia A100 **Nvidia H200**	55 x 55 (5-2-5) 58 x 55 (5-2-5)	GPU + HBM connected with µbumps to Si interposer, connected to laminate build-up substrate
Nvidia Blackwell	70 x 76 (7-2-7)	GPU + HBM connected to RDL interposer with bridges, connected to laminate build-up substrate
AMD AECG (formerly Xilinx) (with FPGA chiplets)	77.5 x 77.5 (8-2-8)	FPGA slices + HBM connected with µbumps to Si interposer, connected to laminate build-up substrate

Source: TechSearch International, Inc.

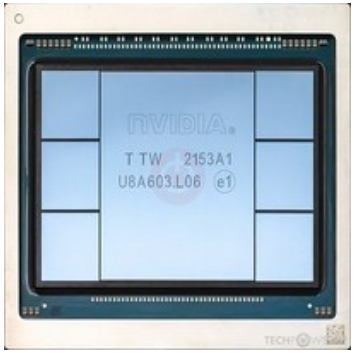
- **Future AI applications with substrate size of 100 mm x 100 mm or larger (120 mm x 120 mm) to support increasing number of HBMs**

© 2025 TechSearch International, Inc.

Nvidia's GPU + HBM with TSMC's CoWoS®-S

- **Nvidia's A100 uses GPU + 6 HBMs**
  - Continues to use TSMC's CoWoS®-S process
  - Si interposer is ~1,600 mm^2
  - Attached to 55mm x 55mm build-up substrate
  - Package has 2,743 balls
- **DeepSeek used 2,000 of Nvidia's H800 SXM5 graphics card with H100 launched on March 21, 2023 and met export requirements**
  - GH100 is 814mm^2 GPU fabricated on TSMC's 5nm process and is packaged using CoWoS®-S
  - 6 HBMs
- **H200s shipping in volume in CoWoS®-S**
  - Features reticle-size GPU with 54µm bump pitch, plus 6 HBM3 stacks

Source: NVIDIA.



techsearchinc.com

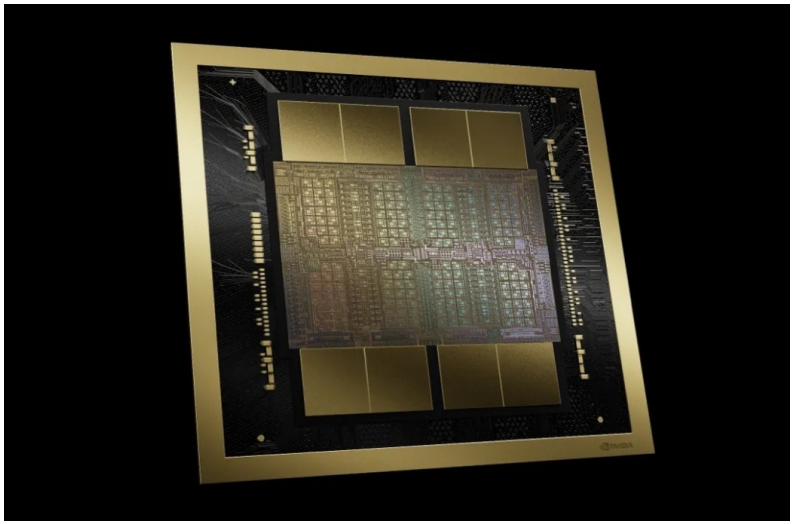
© 2025 TechSearch International, Inc.

TechSearch INTERNATIONAL

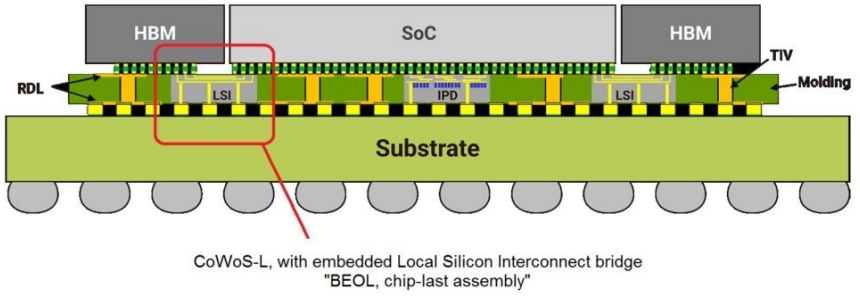
Nvidia GPU + HBM with CoWoS®-L

Blackwell

- **Nvidia's Blackwell shipments started end of 2024**
  - Uses TSMC's CoWoS®-L with RDL interposer with embedded bridge instead of Si interposer
  - 70mm x 76mm package
  - Blackwell B200 dissipates 1,000 W
  - Blackwell B200A comparable to H200 performance, features 1 GPU plus 4 HBM3e stacks and is air-cooled
- **Rubin AI platform will include adoption of HBM4 and will be packaged in CoWoS®-L**
  - Rubin Ultra expected to use 12 stacks of 8-high HBM4E with 5.5-reticle-size interposer attached to 100mm x 100mm substrate



Source: Nvidia.

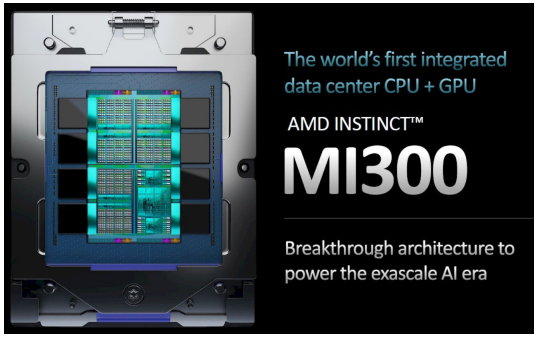


Source: TSMC..

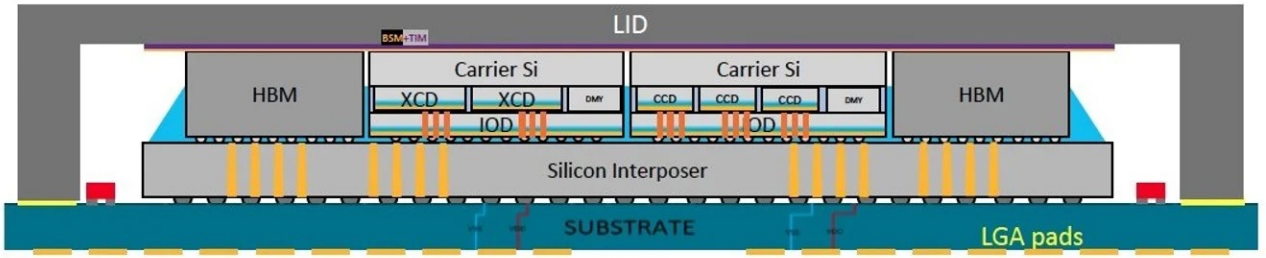
© 2025 TechSearch International, Inc.

AMD Instinct MI300 Series with TSMC's CoWoS®-S

- **Integrates 5nm GPU and CPU logic-to-logic stacking in a 3D hybrid bonding structure using TSMC's SoIC process for chiplets plus 8 stacks of HMB3s**
  - Each HBM is 12-high stack
- **MI300 stacks 3 CPU chiplets (CCDs) and 6 accelerator chiplets (XCDs) on top of 4 IO dies onto a silicon interposer**
- **Uses TSMC's CoWoS®-S process**
  - Si interposer is ~60 mm x 50 mm
  - Package body size of 77.5 mm x 62.5 mm



Source: AMD.



Source: AMD.

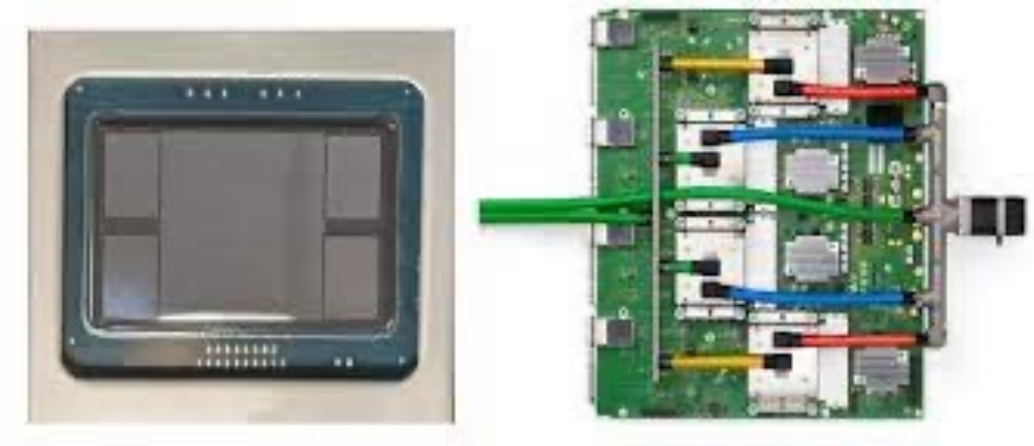
© 2025 TechSearch International, Inc.

AWS Trainium2 with CoWoS®-R

- **AWS Trainium2 for training and inferencing using CoWoS®-R**

© 2025 TechSearch International, Inc.

Google TPUv4 for Machine Learning



Source: Google.

- **Google's Tensor Processing Unit v4 uses a silicon interposer plus HBM2 with 4 per module**
  - Reported performance that exceeds NVIDIA A100

© 2025 TechSearch International, Inc.

Microsoft Maia 100 with CoWoS®-S

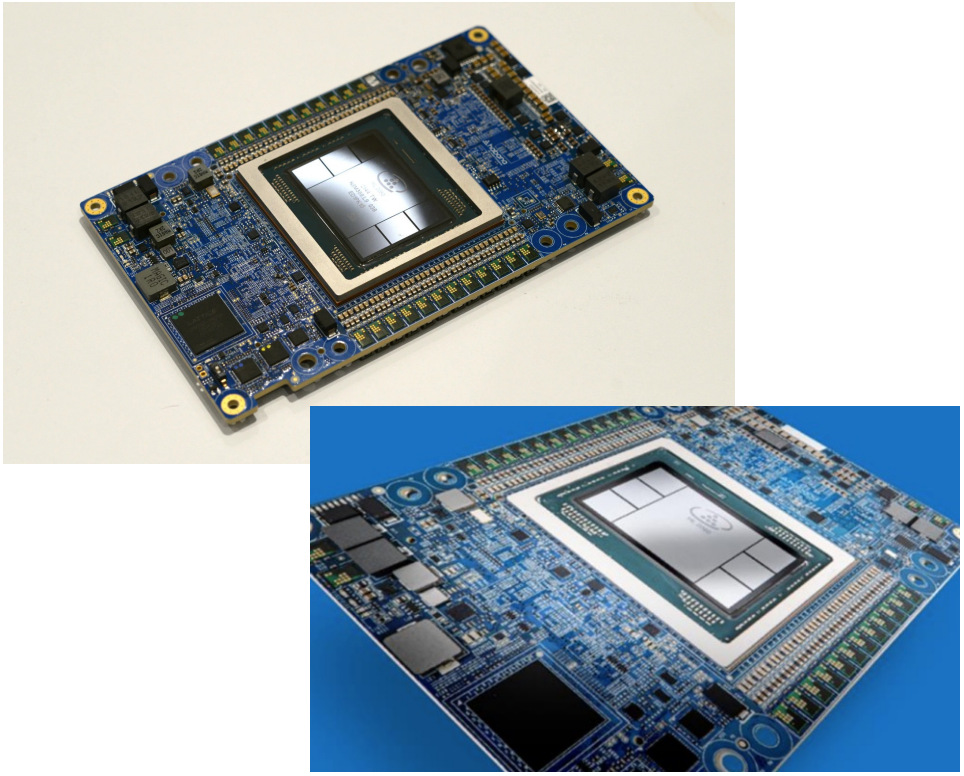
- **Maia 100 is an AI accelerator designed specifically for large-scale workloads deployed in Azure**
  - Features ~820mm^2 chip with on-die SRAM fabricated on TSMC's N5 process and 4 HBM2e stacks
  - Packaged in TSMC's CoWoS®-S

Source: Microsoft.

© 2025 TechSearch International, Inc.

Intel Habana® Gaudi3® Deep Learning Training and Inference Processor

- **Intel purchased Habana in 2019**
- **Gaudi2 chip was fabricated at TSMC on 7nm semiconductor node and packaged by TSMC in CoWoS**
  - Not subject to export control
  - Considered an alternative to Nvidia's A100
- **Gaudi3 is fabricated at TSMC on 5nm node**
  - Used in Amazon AI training

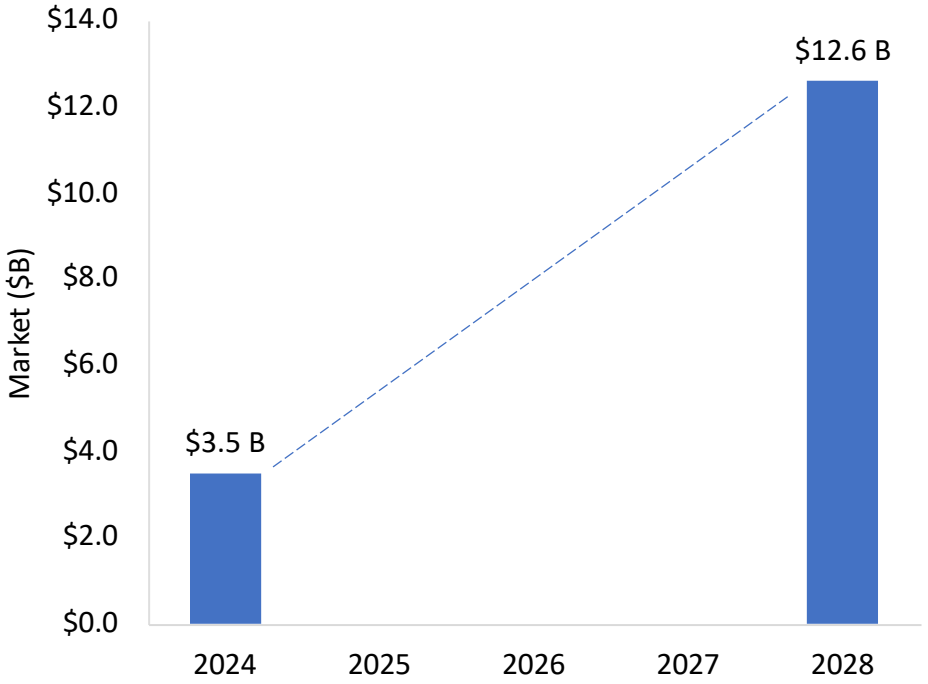


Source: Intel.

© 2025 TechSearch International, Inc.

AI Assembly Revenue Growth

- **Assembly of AI training and inferencing hardware continues to drive advanced packaging revenue growth**
  - AI accounts for 59% of HPC assembly revenue in 2024, increasing to 79% in 2028
- **HPC advanced packaging assembly market is defined as assembly of AI training and inferencing, server, and high-end network switch packages**
- **AI training and inferencing hardware calculated as assembly of logic + HBM on interposer and substrate plus HBM assembly**
  - Edge AI hardware not included
  - CPO not included



Source: TechSearch International, Inc.

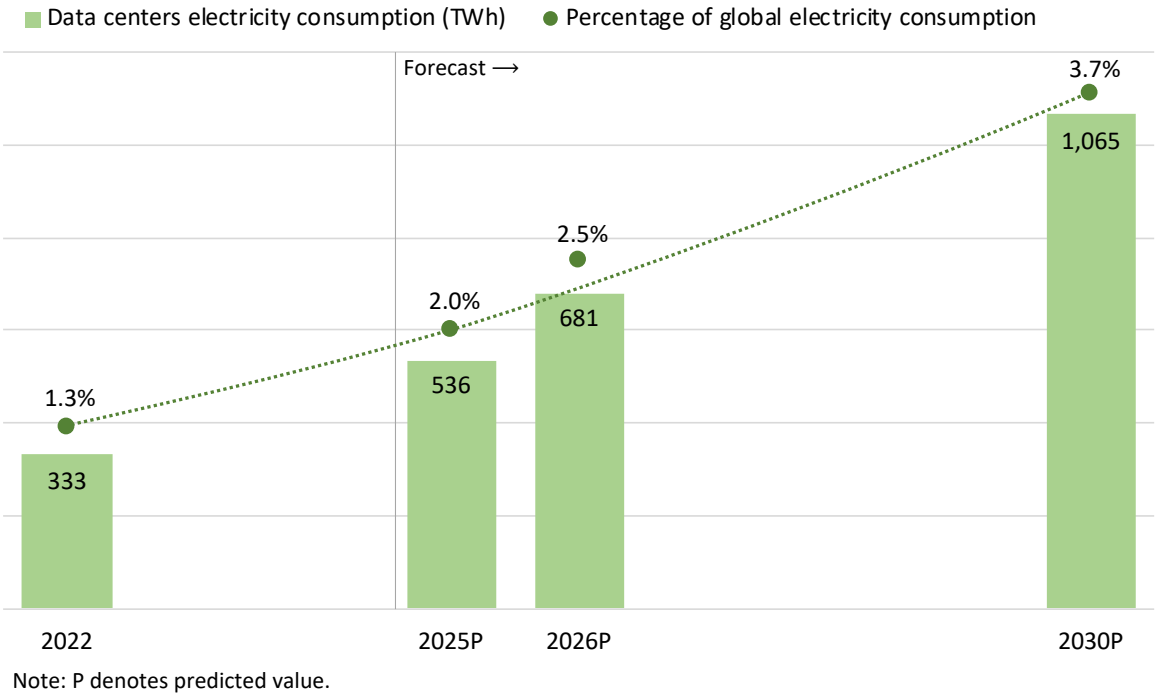
© 2025 TechSearch International, Inc.

Co-Packaged Optics (CPO)

© 2025 TechSearch International, Inc.

TechSearch INTERNATIONAL

Data Center Energy Consumption



Note: P denotes predicted value.

Source: Deloitte.

- **Increased use of AI in data centers resulting in dramatic increase in energy consumption, driving the need for improved I/O power efficiency, especially in hyperscale data centers**

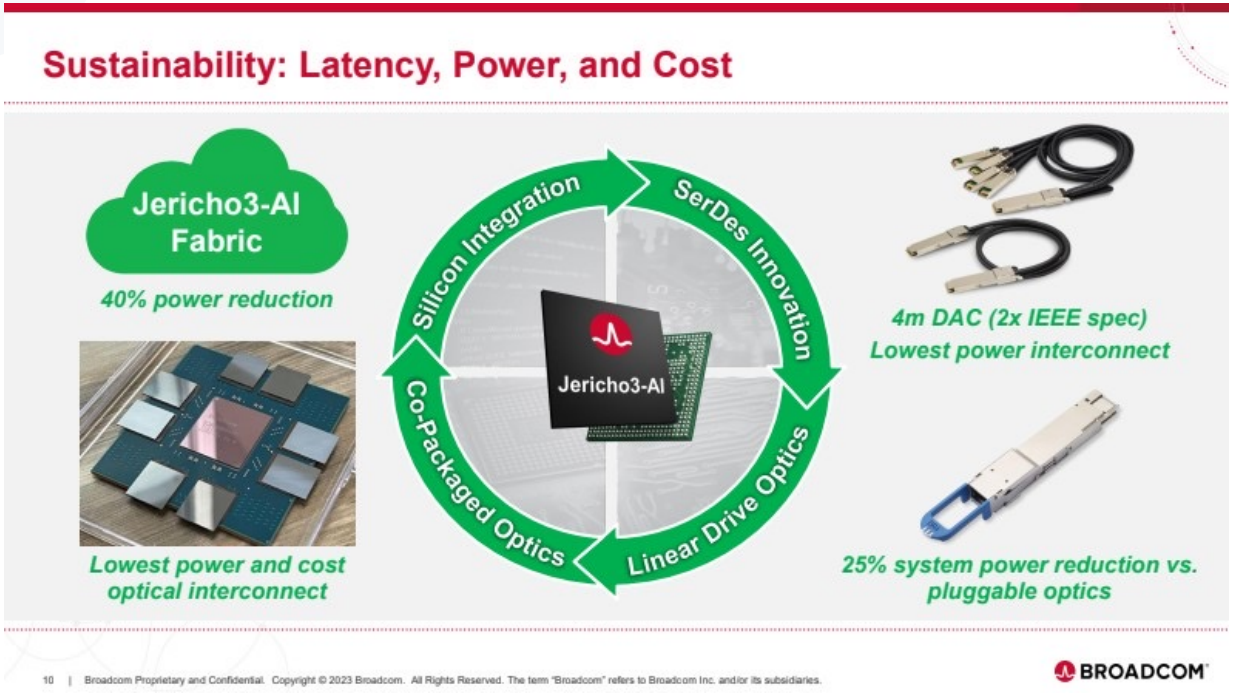
© 2025 TechSearch International, Inc.

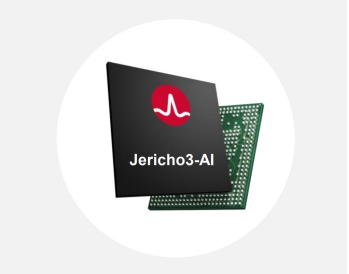
Drivers for Co-Packaged Optics (CPOs)

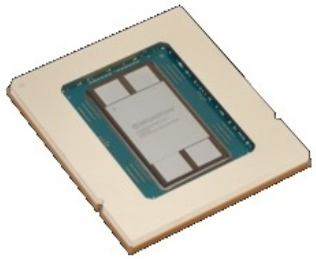
- **Increased network traffic caused by more machine-to-machine traffic as AI and machine learning become greater part of workload**
- **Higher workload increases amount of power consumed by network, which takes power away from other applications (because the datacenter's total power budget is fixed)**
  - Meta says 30% power savings with CPO
- **Partially offset power increases due to higher network bandwidths**
- **Goal is to minimize power of the SerDes interconnect between switch ASIC core and optics**
  - Thermal limits: SerDes very high power + additional for cooling
- **With CPOs link energy (pJ/bit) improves by 30 to 50% (link energy is Power in W that the module dissipates divided by aggregated BW of the chiplet)**
- **Improves system reliability and therefore network availability**
- **Reduces cost (Lower cost per bit by 40%)**

© 2025 TechSearch International, Inc.

Jericho3-AI Machine Learning Ethernet Switch Series







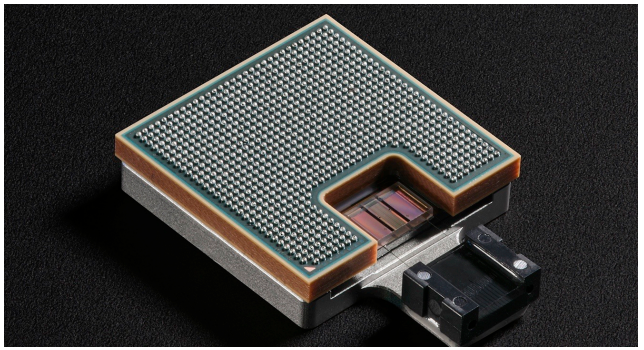
- **AI workloads can be constrained by network latency and bandwidth**
- **Jericho3-AI fabric is designed to lower the time spent in networking during AI training**
- **Broadcom demonstrated a 51.2Tb/s CPO switch that integrates the IC with the PIC**
  - Packaged in LGA

© 2025 TechSearch International, Inc.

CPO Introductions

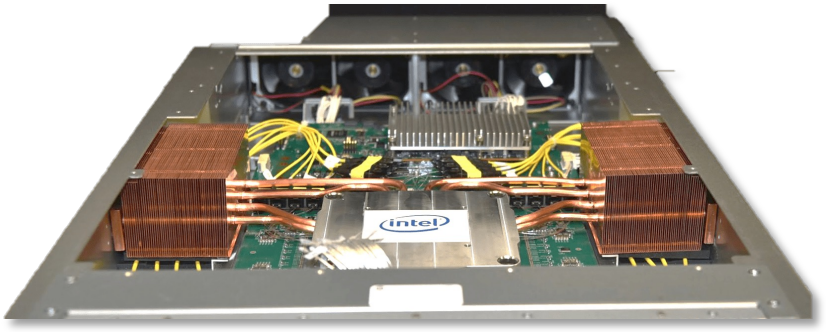
- **IBM introduced a CPO module that is expected to provide an order of magnitude improvement in bandwidth density for data centers**
- **Demonstrated a fully integrated optical compute interconnect (OCI) chiplet co-packaged with an Intel CPU running live data at OFC in 2024 targeted for AI in data centers**
  - Demo with Intel CPU, but can also be integrated with next-gen CPUs, GPUs, IPUs, and other SoCs
- **Nvidia expected to announce a CPO product this month**

IBM CPO



Source: IBM.

Intel CPO



Source: Intel.

© 2025 TechSearch International, Inc.

Edge AI

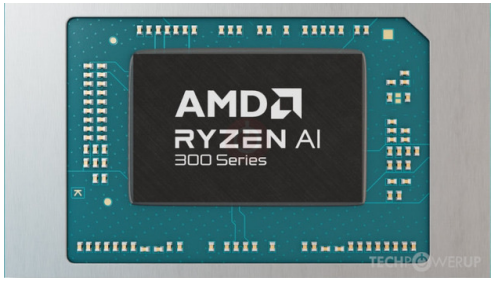
© 2025 TechSearch International, Inc.

TechSearch INTERNATIONAL

Edge AI

- **PCs**
  - AI PC defined using Microsoft Copilot's benchmark of 40 Tera (trillion) Operations per Second (TOPS)
  - Potential for >100 million AI PCs in 2025
- **Smartphones**
  - IDC defines next-generation AI smartphones as devices capable of running on-device Generative AI (GenAI) models more quickly and efficiently leveraging a neural processing unit (NPU) with 30 TOPS
  - Almost all high-end phones AI capable

FC-BGA

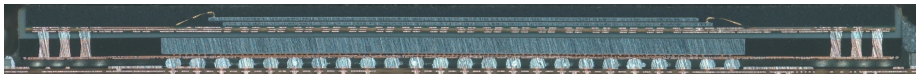


Source: AMD.

FC-BGA

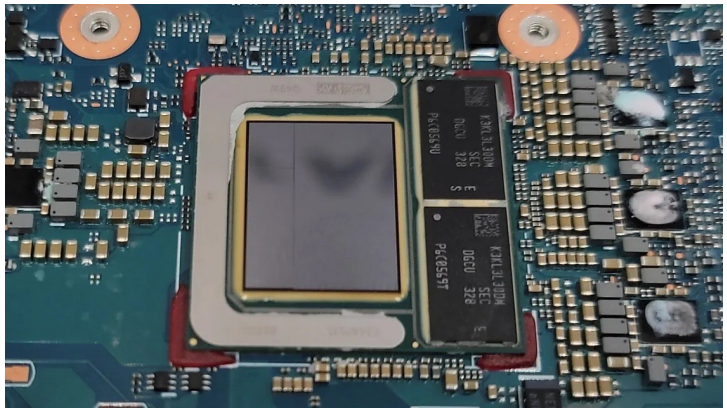
Source: Qualcomm.

Apple AP in TSMC's InFO PoP



Source: TechSearch International, Inc. (Comet Yxlon provided equipment for x-ray).

3D Foveros with µbumps



Source: Intel.

© 2025 TechSearch International, Inc.

Edge AI (con.)

- **Automotive**
  - Chiplet designs such as Renesas 5th Gen R-Car processor SOC designed on 3nm semiconductor node for ADAS, in-vehicle infotainment, and gateway applications
- **Robotics**
  - Adopting AI for training (Nvidia announced Cosmos foundation model platform)
  - Goldman Sachs forecasts humanoid robot shipments will exceed 250,000 units by 2030
- **Home appliances**
- **Smart glasses**
  - Meta Ray-Ban sold >2 million Ai-equipped smart glasses
  - Shipments projected >20 million units by 2028

Renesas R-Car

Source: Renesas.

3D Foveros with µbumps

Source: Intel.

Source: Google.

© 2025 TechSearch International, Inc.

AI Infrastructure

© 2025 TechSearch International, Inc.

TechSearch INTERNATIONAL

Additional Industry Segments Benefit from AI Growth

- **Test equipment suppliers**
  - Trend in integrating automated test equipment (ATE) and system level test (SLT)
  - Advantest raised its outlook for test equipment 40% for the year in anticipation of elevated spending for test equipment used for AI
- **PCB makers and related supply chain such as CCL suppliers and resin makers**
  - AI servers with high-layer count PCBs
  - Need high-frequency and high-speed transmission materials
- **Power modules**
  - Growth in embedded die
- **Capacitors**
  - Murata predicts demand for MLCCs to double this year
  - Conventional server has 2,000 caps per motherboard, AI server anticipated to need 200,000 capacitors
- **Storage devices**
  - Specifically, NAND Flash

© 2025 TechSearch International, Inc.

Conclusions

- **AI growth will continue**
  - Demand for AI training and inferencing continues to grow, driving industry advanced packaging revenue growth, even with small unit volumes
  - Larger volumes with edge applications
- **Many different packages for AI applications**
  - Silicon interposer attached to build-up substrate
  - RDL interposer attached to build-up substrate (with or without bridges)
  - 3D with µbumps
  - Flip chip BGA
  - Fan-out WLP
- **Growth of AI benefits many segments of the industry**

© 2025 TechSearch International, Inc.

Thank you!

TechSearch International, Inc.
4801 Spicewood Springs Road, Suite 150
Austin, Texas 78759 USA
+1.512.372.8887
tsi@techsearchinc.com

RELEVANT, ACCURATE, TIMELY

© 2025 TechSearch International, Inc.