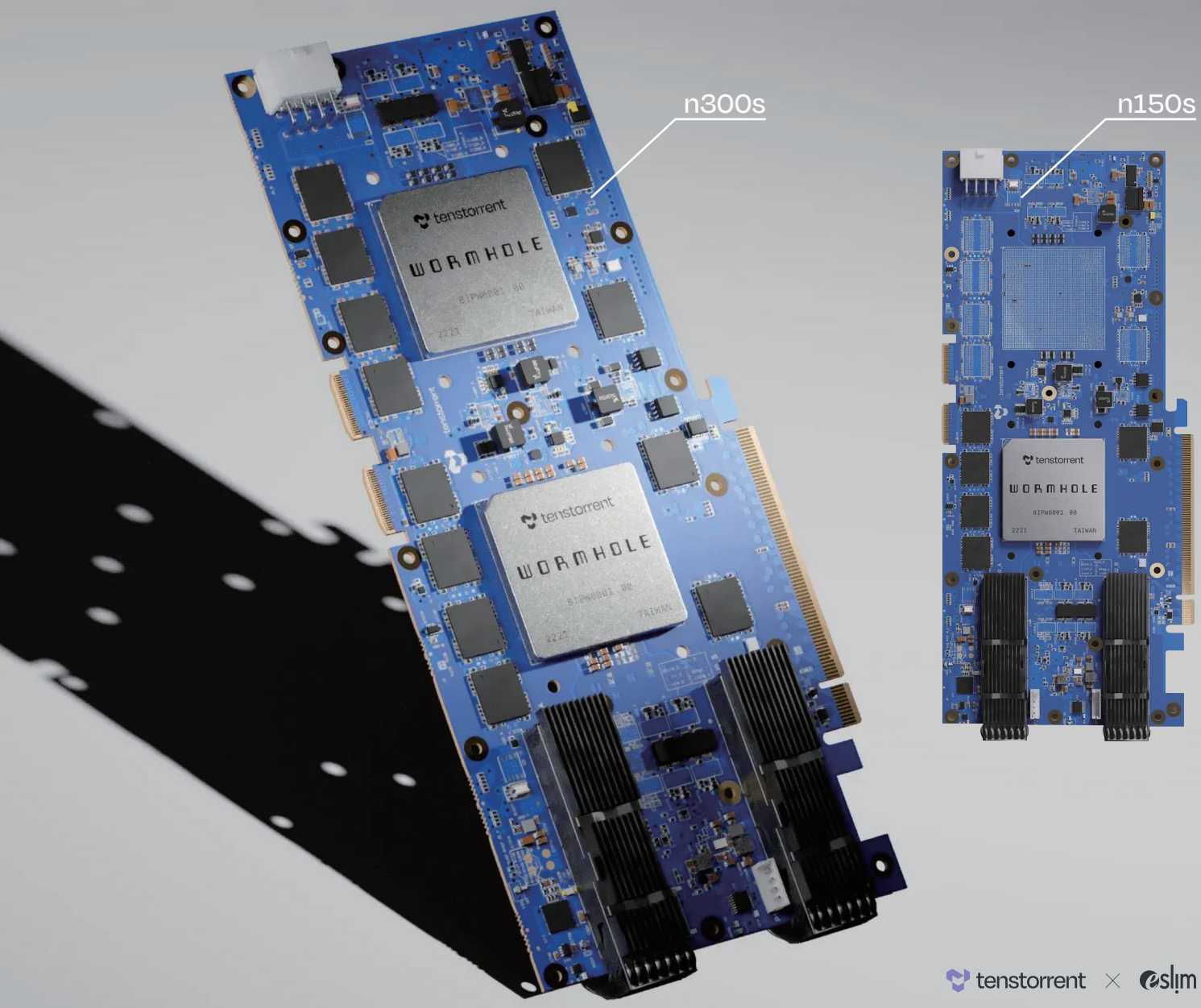
TT-Wormhole™

RISC-V 아키텍처 기반, 고성능-저전력 AI 가속기



AI 연산의 패러다임, 더 높고 새로운 차원으로

TT-Wormhole™은 차세대 AI 가속기로, **TT-Forge™**와 **TT-Metalium™** 같은 소프트웨어 스택과 함께 작동하여 고성능 AI 및 범용 연산을 지원하는 하드웨어 플랫폼입니다.

TT-Wormhole™은 **대규모 병렬 처리와 고속 데이터 이동에 최적화**되어 있어, 추론과 훈련 워크로드 모두에 효과적으로 대응할 수 있도록 설계되었습니다.

텐스토렌트
한국공식총판

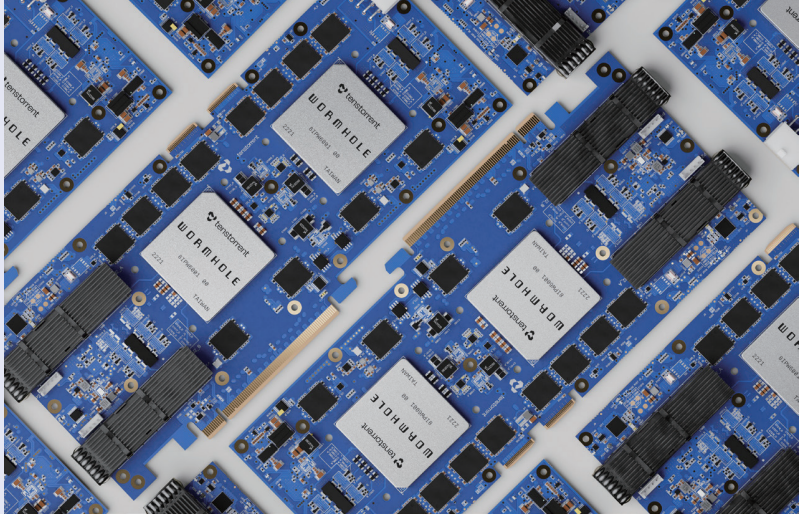
이슬림코리아
www.eslim.co.kr

간편하고 효율적인 IT · AI 인프라는 역시 이슬림코리아! 지금 바로 이슬림코리아 홈페이지에 접속해 다양한 정보를 확인해 보세요!

tenstorrent

Wormhole™ Tensix 프로세서

확장 가능 • 커스터마이징 가능 • 경제적



Wormhole™은 Tenstorrent의 1세대 Tensix 프로세서 기반 AI 가속기로, AI 및 고성능 컴퓨팅(HPC) 워크로드를 위해 설계되었습니다. Tensix Core 아키텍처를 기반으로 구축되며, 각각의 코어는 컴퓨팅 장치, 네트워크-온-칩(NoC), 로컬 캐시, 'Baby RISC-V' 코어를 포함하여 효율적인 데이터 이동과 고속 연산을 지원합니다. Wormhole™은 AI 학습 및 추론에 최적화되어 있으며, 뛰어난 병렬 처리 성능과 확장성을 제공합니다. 또한, 개발자는 Tenstorrent의 오픈 소스 SDK인 TT-Metalium™(로우 레벨) 및 TT-Forge™(중간 레벨 컴파일러)를 활용하여 모델을 최적화된 실행 그래프로 변환하고, 하드웨어 자원을 효율적으로 활용할 수 있습니다.. Tenstorrent는 동급 가격대의 AI 및 HPC 가속기와 비교하여 더 뛰어난 메모리 용량과 성능을 제공하며, AI 및 데이터 연산을 위한 강력한 솔루션을 제공합니다.

유연한 설계

강력함 :
n150은 160W의 전력으로 최대 262 TFLOPs(FP8), n300은 300W의 전력으로 최대 466 TFLOPs(FP8)의 연산 성능을 제공하며, 다양한 고성능 AI 워크로드에 대응할 수 있습니다.

확장성 :
워크스테이션 및 랙마운트 서버 환경에서 Ethernet 기반의 멀티칩 메시 구조를 통해 연산 자원을 유연하게 확장할 수 있습니다.

오픈 소스 :
TT-Forge™와 TT-Metalium™ SDK를 통해 고수준부터 로우레벨까지의 개발을 지원하며, 모델 실행 최적화와 시스템 통합이 가능합니다.

커스터마이징 :
연산 속도를 우선시할지, 정확도를 중시할지를 선택적으로 조정할 수 있어, 다양한 AI 애플리케이션에 최적화된 성능을 구현할 수 있습니다.

유연성 :
FP8, FP16, BFLOAT16, FP32(출력 전용), BLOCKFP2/4/8, INT8/32(출력 전용), UINT8, TF32, VTF19, VFP32 등 폭넓은 연산 포맷을 지원합니다.

멀티 유저 친화적 :
Wormhole 기반 시스템에서는 카드와 ASIC을 개별 타겟팅할 수 있어, 여러 사용자가 동시에 병렬 처리를 수행하는 환경에 적합합니다.

Wormhole™ Tensix 프로세서

** 텐스토렌트 Wormhole™ 기반 카드에만 연결할 수 있습니다.*

Card	n150s / n150d	n300s / n300d
Wormhole™	1	2
Tensix Cores	72	128 (64 per ASIC)
AI Clock	1 GHz	1 GHz
SRAM	108MB	192MB (96MB per ASIC)
Memory	12GB GDDR6	24GB GDDR6
Memory Speed	12 GT/sec	12 GT/sec
Memory Bandwidth	288 GB/sec	576 GB/sec
TFLOPS (FP8)	262	466
TFLOPS (FP16)	74	131
TFLOPS (BLOCKFP8)	148	262
TBP(Total Board Power)	160W	300W
Connectivity	2x Warp 100 Bridge, 2x QSFP-DD 200G (Active)*	2x Warp 100 Bridge, 2x QSFP-DD 200G (Active)*
Internal Chip-to-Chip	N/A	400GbE
System Interface	PCI Express 4.0 x16	PCI Express 4.0 x16
Cooling	Passive(Cooling Kit) / Active (Axial Fan)	Passive(Cooling Kit) / Active (Axial Fan)
FormFactor	Dual Slot, FHFL / 2.5-Slot, FHFL	Dual Slot, FHFL / 2.5-Slot, FHFL