

IBM, IBM Z AI

- IBM II IBM , AI AI
- 5nm

2024 8 27 , – IBM ‘ (Hot Chips) 2024’ IBM (Telum) II IBM (Spyre Accelerator) . IBM Z (AI) AI AI .

(Large Language Model, LLM) AI . 8 [1] AI 75% 2026 2022 . IBM AI .

.

- IBM II : IBM Z IBM 1 () , 40% AI , (DPU) LLM .
- IO : II (DPU) IO . DPU .
- IBM : II AI . II AI LLM AI . AI 2024 IBM . 75 PCIe . PCIe

(Tina Tarquinio) IBM Z (LinuxONE) “IBM AI ” , “ II , , IBM Z LLM AI .

II IBM IBM 5nm . AI . AI . , AI . :

. : LLM AI .

. :

. AI : , , .

II IBM Z IBM , IBM Z . IBM .

[]

II : 5.5GHz 8 36MB L2 , 40% (360MB) . (drawer) 2.88GB 4 40% . AI (Latency) AI , 4 . IO DPU II . 50% IO . IBM Z AI .

: AI AI . IO 8 1TB , 75W AI .
AI int4, int8, fp8 fp16 32 .

IBM

IBM AI, 175 , , , . , 3,000 IBI
. IBM AI, , . , , , IBM .
www.ibm.com .

* IBM , .

[1] : Morgan Stanley 2024 8

###