



The science behind the report:

# Dell PowerEdge XE9680 servers with AMD Instinct MI300X Accelerators: the power to host GenAI with Llama 3.1 405B LLMs

This document describes what we tested, how we tested, and what we found. To learn how these facts translate into real-world benefits, read the report [Dell PowerEdge XE9680 servers with AMD Instinct MI300X Accelerators: the power to host GenAI with Llama 3.1 405B LLMs](#).

We concluded our hands-on testing on March 31, 2025. During testing, we determined the appropriate hardware and software configurations and applied updates as they became available. The results in this report reflect configurations that we finalized on March 22, 2025 or earlier. Unavoidably, these configurations may not represent the latest versions available when this report appears.

## Our results

To learn more about how we have calculated the wins in this report, go to <http://facts.pt/calculating-and-highlighting-wins>. Unless we state otherwise, we have followed the rules and principles we outline in that document.

Table 1: Results of our PTChatterly testing.

	Dell PowerEdge XE9680 4 x AMD Instinct MI300X Accelerators	Dell PowerEdge XE9680 8 x AMD Instinct MI300X Accelerators
Results		
Max users reached	72	136
Cut off on question response time (s)	30	30
Threshold percentage	95%	95%
Median threshold (s)	15	15
CPU AVG %UTIL @ max users	3.3	6.1
GPU AVG %UTIL @ max users	97.2	97.2
Idle power (W)	4,802	4,998
Active power (W)	6,459.3	8,029.3

	Dell PowerEdge XE9680 4 x AMD Instinct MI300X Accelerators	Dell PowerEdge XE9680 8 x AMD Instinct MI300X Accelerators
Test parameters		
LLM version	rocm/vllm:rocm6.3.1_mi300_ubuntu22.04_py3.12_vllm_0.6.6	rocm/vllm:rocm6.3.1_mi300_ubuntu22.04_py3.12_vllm_0.6.6
LLM model	amd/Llama-3.1-405B-Instruct-FP8-KV	amd/Llama-3.1-405B-Instruct-FP8-KV
LLM instances	1	2
Corpus	airbnb	airbnb
Corpus # of listings	487,974	487,974
Corpus total text size (MB)	1,952.69	1,952.69
# of questions in conversation	1-9	1-9
Think time (s)	2.5	2.5
Think time range (s)	1	1
Ladder delay(s) (divided by number of users for each run)	10	10
User delay basis (s)	30	30
Max user delay (s)	0.1	0.1
Maximum # of tokens per prediction	128	128
Prompt context size	8,192	8,192
Threads (CPU only)	Default	Default
Full options for LLM service	<pre>"VLLM_USE_TRITON_FLASH_ATTN=0 --swap-space 16 --disable-log-requests --dtype float16 --quantization fp8 --kv-cache-dtype fp8 --tensor-parallel-size 4 --gpu-memory-utilization 0.9 --max-model-len 8192 --max-num-batched-tokens 8192 --num-scheduler-steps 1 --max-seq-len-to-capture 256 --max-num-seqs 256 --distributed-executor-backend mp --enable-chunked-prefill --enable-prefix-caching"</pre>	<pre>"VLLM_USE_TRITON_FLASH_ATTN=0 --swap-space 16 --disable-log-requests --dtype float16 --quantization fp8 --kv-cache-dtype fp8 --tensor-parallel-size 4 --gpu-memory-utilization 0.9 --max-model-len 8192 --max-num-batched-tokens 8192 --num-scheduler-steps 1 --max-seq-len-to-capture 256 --max-num-seqs 256 --distributed-executor-backend mp --enable-chunked-prefill --enable-prefix-caching"</pre>
Benchmark mode	Timed Ladder	Timed Ladder
Begin run with # users	1	1
Increment by # users	1	1
Broker endpoint	http://100.67.190.137:8072/prompts	http://100.67.190.137:8072/prompts
Broker version	v2025.03.2.10 (22 March 2025)	v2025.03.2.10 (22 March 2025)
Client version	v2025.03.2.09 (18 March 2025)	v2025.03.2.10 (22 March 2025)

## Cost analysis

In Tables 2 through 7, we detail the various costs and assumptions we used to create our example 5-year TCO for 6x Dell PowerEdge XE9680 servers each with eight AMD Instinct MI300X Accelerators. As with any TCO calculation, your expected costs will vary depending on several factors.

### Total systems required for equivalent performance

Table 2: The results of our PTChatterly testing extrapolated to the number of users a rack of six servers could support. Source: Principled Technologies.

	6 x Dell PowerEdge XE9680 with AMD Instinct MI300X Accelerators	Notes
Servers		
Total number of systems in one rack	6	Number of servers that fit in a 60kW rack
Total number of users in one rack	816	136 simultaneous AI chatbot users per server times 6 servers

### Hardware costs

Table 3: Estimated hardware costs for 6x Dell PowerEdge XE9680 with AMD Instinct MI300X Accelerators. Source: Principled Technologies.

	6x Dell PowerEdge XE9680 with AMD Instinct MI300X Accelerators	Notes
Configuration details		
System	1x PowerEdge XE9680	
CPU	2x Intel Xeon Platinum 8468	
Memory	32x 64GB DDR5	
OCP NIC	1x Broadcom Gigabit Ethernet BCM5720	
Power supply	6x 2800W PSU	
NVMe	8x 3.84 NVMe U.2	
Boss	2x 480GB M.2	
GPU	AMD Instinct MI300X 8-way	
<b>Total cost for one system</b>	<b>\$1,213,371.49</b>	
<b>Total cost for six systems (one rack)</b>	<b>\$7,280,228.94</b>	

## Power and cooling costs

Table 4: Power and cooling cost estimates over five years for 6x Dell PowerEdge XE9680 with AMD Instinct MI300X Accelerators. Source: Principled Technologies.

	6x Dell PowerEdge XE9680 with AMD Instinct MI300X Accelerators	Notes
Cost per kWh	0.1701	Avg price of electricity Nov 2024 <a href="https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=table_es1a">https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=table_es1a</a>
Hours in year	8,760	
Percentage time under load/active (remainder idle)	0.50	Assume 50% active (12 hours per day), but this can be set to any percentage of total operating time we feel comfortable assuming the servers would be under full load.
Power usage per server (watts-active)	8,029	
Power usage per server (watts-idle)	4,998	
Typical watts	6513.5	Calculation: (% time active * watts-active) + (% time idle * watts-idle)
Annual kWh per server	57058.26	
Total annual energy cost per server	\$9,705.61	
<b>Total 5y energy cost per server</b>	<b>\$48,528.05</b>	
<b>Total 5y cost for all required systems</b>	<b>\$291,168.30</b>	

## Datacenter space costs

Table 5: Data center space cost estimates over five years for 6x Dell PowerEdge XE9680 with AMD Instinct MI300X Accelerators. Source: Principled Technologies.

	6x Dell PowerEdge XE9680 with AMD Instinct MI300X Accelerators	Notes
Rack units per server (u)	6	
Annual data center costs per rack (42u)	\$2,000.00	PT estimate. Could vary \$1,000 - \$3,000+ depending on location, bandwidth, infra, and other factors. See <a href="https://cyfuture.cloud/kb/colocation/how-much-does-renting-rack-space-cost-key-factors-to-consider">https://cyfuture.cloud/kb/colocation/how-much-does-renting-rack-space-cost-key-factors-to-consider</a> . Including infrastructure requirements, a full cabinet is assumed.
<b>Annual cost for all required systems</b>	<b>\$2,000.00</b>	
<b>Total 5y cost for all required systems</b>	<b>\$10,000.00</b>	

## Maintenance and administration costs

Table 6: Maintenance and administration cost estimates over five years for 6x Dell PowerEdge XE9680 with AMD Instinct MI300X Accelerators. Source: Principled Technologies.

	6x Dell PowerEdge XE9680 with AMD Instinct MI300X Accelerators	Notes
Number of servers per IT admin	100	PT estimate. Could vary significantly based on management/infrastructure factors.
Number admins needed for all required systems	0.06	Total required systems divided by number of servers per IT admin.
Average salary of an administrator	\$100,580.00	Average for network and computer systems administrator, BLS May 2023 <a href="https://www.bls.gov/oes/current/oes_nat.htm">https://www.bls.gov/oes/current/oes_nat.htm</a> .
Burden rate	0.2962	Burden rate for private industry workers, BLS Sept 2024 <a href="https://www.bls.gov/news.release/ecec.nr0.htm">https://www.bls.gov/news.release/ecec.nr0.htm</a>
Average burdened salary	\$130,368.90	Average salary * (1 + burden rate)
Annual administration cost	\$7,822.13	
<b>Total 5y administration cost</b>	<b>\$39,110.67</b>	

## Operational costs

Table 7: Operational cost estimates over five years for 6x Dell PowerEdge XE9680 with AMD Instinct MI300X Accelerators. Source: Principled Technologies.

	6 x Dell PowerEdge XE9680 with AMD Instinct MI300X Accelerators	Notes
Number of systems required	6	
Total system cost	\$7,280,228.94	
Total 5y power cost	\$291,168.30	
Total 5y datacenter space cost	\$10,000.00	
Total 5y maintenance cost	\$39,110.67	
<b>Total 5y costs</b>	<b>\$7,620,507.91</b>	

## System configuration information

Table 8: Detailed information on the systems we tested.

System configuration information	Dell PowerEdge XE9680 server	Dell PowerEdge R660 client
BIOS name and version	Dell 2.4.4	Dell 2.4.4
Non-default BIOS settings	N/A	N/A
Operating system name and version/build number	Ubuntu 22.04.5 LTS Kernel version 5.15.0-133-generic	Ubuntu 22.04.5 LTS Kernel version 5.15.0-133-generic
Date of last OS updates/patches applied	2/24/25	2/24/25
Power management policy	Performance	Performance
Processor		
Number of processors	2	1
Vendor and model	Intel® Xeon® Platinum 8468	Intel Xeon Gold 6448Y
Core count (per processor)	48	32
Core frequency (GHz)	2.1	2.1
Max turbo frequency (GHz)	3.8	4.1
Stepping	6	8
Memory module(s)		
Total memory in system (GB)	2,048	512
Number of memory modules	32	16
Vendor and model	Samsung® M321R8GA0BB0-CQKZJ	Hynix® HMC88AEBRA107N
Size (GB)	64	32
Type	PC5-38400	PC5-38400
Speed (MT/s)	4,800	4,800
Speed running in the server (MT/s)	4,800	4,800
Storage controller		
Vendor and model	N/A	PERC H965i Front (Embedded)
Cache size (GB)	N/A	8GB
Firmware version	N/A	8.8.0.0.18-31
Driver version	N/A	8.0.0.69.0
Local storage		
Number of drives	1	4
Drive vendor and model	Dell NVMe® PM1733a RI 3.84TB	Samsung PM1645a MZILT800HBHQ0D3
Drive size (GB)	2,980	800
Drive information (speed, interface, type)	PCIe gen4x4, NVMe, SSD	12Gbps, SAS, SSD
Purpose	OS, application	OS, application

System configuration information	Dell PowerEdge XE9680 server	Dell PowerEdge R660 client
Network adapter		
Vendor and model	Broadcom® BCM5720 Gigabit Ethernet	Broadcom BCM5720 Gigabit Ethernet
Number and type of ports	2 x 1GbE	2 x 1GbE
Driver version	tg3 5.15.0-133-generic	tg3 5.15.0-133-generic
Accelerators		
Vendor and model	AMD® Instinct™ MI300X	N/A
Number	8	N/A
Memory type	HBM3	N/A
Memory size	192GB	N/A
Memory bandwidth	5.3 TB/s	N/A
Power cap limit	750W	N/A
Connection	Infinity Fabric™, PCIe gen5, 16 lanes	N/A
Driver	6.3.1.60301	N/A
Cooling fans		
Vendor and model	Dell Gold	Dell Gold
Number of cooling fans	32	16
Power supplies		
Vendor and model	Dell 0FX9WJA01	Dell 07DWCYA01
Number of power supplies	6	2
Wattage of each (W)	2,800	1,400

# How we tested

## Software versions

- OS: Ubuntu Server 22.04.5 LTS (kernel version 5.15.0-133-generic)
- ROCm driver: 6.3.1.60301-48~22.04
- Docker: docker-ce 5:28.0.0-1~ubuntu.22.04~jammy
- HAproxy: 2.4.24-0ubuntu0.22.04.1
- VectorDB: qdrant/qdrant:v1.13.4
- VectorDB model: sentence-transformers/msmarco-distilbert-cos-v5
- VectorDB precision: FP32
- Embedding server: michaelf34/infinity:0.0.67
- Embedding server model: sentence-transformers/msmarco-distilbert-cos-v5
- Embedding server precision: FP32
- Go Lang version: 1.23.1
- PTChatterly client: v2025.03.2.09 (18 March 2025)
- PTChatterly broker: v2025.03.2.10 (22 March 2025)
- LLM server (AMD GPU): rocm/vllm:rocm6.3.1\_mi300\_ubuntu22.04\_py3.12\_vllm\_0.6.6
- LLM model (AMD GPU): amd/Llama-3.1-405B-Instruct-FP8-KV
- LLM model precision: FP8
- Nmon: 16q

## Configuring the system under test (SUT)

1. Install Ubuntu Server 22.04.5 LTS, making sure that sshd is included and running.
2. Configure Ubuntu OS following the instructions in the section Configuring Ubuntu 22.04.
3. Install Docker following the instructions in the section Installing Docker on Ubuntu.
4. Install ROCm drivers following the instructions in the section Installing amdgpu drivers and ROCm.
5. Install the nmon resource monitoring tool on the SUT. You can copy the resulting binary file(s) to the remaining systems:

```
# Remove any preinstalled versions of nmon
sudo apt remove -y nmon
# Build requirements:
# ncurses-dev
# gcc, make, and wget (already installed)
sudo apt install ncurses-dev
# Download source files
mkdir nmon_build
cd nmon_build
wget 'https://sourceforge.net/projects/nmon/files/lmon16q.c/download' -O lmon.c
wget 'https://sourceforge.net/projects/nmon/files/makefile/download' -O Makefile
make
sudo install nmon_X86_Ubuntu22_16q /usr/local/sbin/nmon
```

6. Install nmonchart, the nmon parser and chart creator, on the client system:

```
# Set up nmonchart in a convenient directory on the client system
# The script uses the Korn shell
sudo apt install ksh
wget 'https://sourceforge.net/projects/nmon/files/nmonchart42.tar/download' -O nmonchart42.tar
tar -xvf nmonchart42.tar ./nmonchart
# Install nmonchat
sudo install nmonchart /usr/local/sbin/nmonchart
```

## Ingesting Airbnb data into the vector database

1. Create a new directory on the SUT:

```
mkdir ingest
cd ingest
```

2. Create and activate a new virtual Python environment:

```
sudo apt install python3-venv
python3 -m venv .ingest
. .ingest/bin/activate
```

3. Copy the following files to the directory above:

- Property listings file dataset: AirbnbProps-20240830.json.gz
- Python ingestion script: ingestAirBnB.py

4. Create symbolic link to property listings file dataset:

```
ln -s AirbnbProps-20240830.json.gz AirbnbProps.json.gz
```

5. Add Python packages:

```
pip3 install -U torch --index-url https://download.pytorch.org/whl/rocm6.2.4
pip3 install -U sentence-transformers
pip3 install -U qdrant-client
```

## Configuring the associated client / test harness server

1. Install Ubuntu Server 22.04.5 LTS, making sure that `sshd` is included and running.
2. Configure Ubuntu OS following the instructions in Appendix 1.
3. Install Docker following the instructions in Appendix 2.
4. Install HAProxy:

```
sudo apt update
sudo apt install haproxy
sudo systemctl disable haproxy
```

5. Install the PTChatterly broker and client.

## Configuring Ubuntu 22.04

After installation, perform these configuration steps. We assume the login for the non-root user is `ptuser`.

1. Enable password-less sudo:

```
echo "$USER ALL=(ALL:ALL) NOPASSWD: ALL" | sudo tee /etc/sudoers.d/$USER
sudo chmod 640 /etc/sudoers.d/$USER
```

2. Set time zone:

```
sudo timedatectl set-timezone America/New_York
```

3. Extend the root LVM filesystem, if necessary:

```
sudo lvextend -r -l +100%FREE /dev/ubuntu-vg/ubuntu-lv
```

4. Disable unattended package updates:

```
sudo systemctl stop unattended-upgrades.service
sudo systemctl disable unattended-upgrades.service
```

5. Modify the value of the unattended-upgrade variable to 0 in file `/etc/apt/apt.conf.d/20auto-upgrades`:

```
APT::Periodic::Update-Package-Lists "1";
APT::Periodic::Unattended-Upgrade "0";
```

6. Install the latest updates:

```
sudo apt update
sudo apt upgrade
```

7. Install standard Ubuntu packages you will need in subsequent steps:

```
sudo apt install ca-certificates curl wget lsb-release sysstat smartmontools vim nmon numactl
```

8. Reboot the system:

```
sudo shutdown -r now
```

## Installing Docker on Ubuntu 22.04

1. Remove any previous Docker packages:

```
for pkg in docker.io docker-doc docker-compose docker-compose-v2 \
  podman-docker containerd runc; do \
  sudo apt remove $pkg
done
```

2. Add Docker's official GPG key:

```
sudo install -m 0755 -d /etc/apt/keyrings
sudo curl -fsSL https://download.docker.com/linux/ubuntu/gpg -o /etc/apt/keyrings/docker.asc
sudo chmod a+r /etc/apt/keyrings/docker.asc
```

3. Add the Docker repository to the system:

```
echo \  
  "deb [arch=$(dpkg --print-architecture) signed-by=/etc/apt/keyrings/docker.asc] https://download.  
docker.com/linux/ubuntu \  
  $(. /etc/os-release && echo "$VERSION_CODENAME") stable" | \  
  sudo tee /etc/apt/sources.list.d/docker.list > /dev/null  
sudo apt update
```

4. Install Docker CE:

```
sudo apt update  
sudo apt install docker-ce docker-ce-cli containerd.io docker-buildx-plugin docker-compose-plugin
```

5. Add current user to Docker permissions group, and add docker group to the user's current session:

```
sudo usermod -aG docker $USER  
newgrp docker
```

6. Confirm Docker function under non-root user:

```
docker run --rm hello-world
```

7. Enable Docker services:

```
sudo systemctl enable --now docker.service  
sudo systemctl enable --now containerd.service
```

## Installing amdgpu drivers and ROCm

These instructions assume Ubuntu 22.04.5.

1. Install prerequisite packages:

```
sudo apt install "linux-headers-$(uname -r)" "linux-modules-extra-$(uname -r)"  
sudo apt install python3-setuptools python3-wheel libpython3.10
```

2. Configure permissions for GPU access:

```
sudo usermod -a -G video,render $LOGNAME
```

3. Download the amdgpu-install package to add the amdgpu repository:

```
wget https://repo.radeon.com/amdgpu-install/6.3.1/ubuntu/jammy/amdgpu-install_6.3.60301-1_all.deb  
sudo apt install ./amdgpu-install_6.3.60301-1_all.deb
```

4. Install amdgpu drivers:

```
sudo amdgpu-install --usecase=dkms
```

5. Install ROCm:

```
sudo amdgpu-install --usecase=rocm
```

6. Reboot.
7. Validate GPU functionality (Optional):

```
sudo docker run --rm -it --device=/dev/kfd --device=/dev/dri --security-opt seccomp=unconfined \
--group-add video rocm/rocm-terminal sudo rocminfo
```

## LLM parameters:

```
# Adjust --tensor-parallel-size to the number of GPUs you wish to allocate to your vLLM instance
VLLM_USE_TRITON_FLASH_ATTN=0
--swap-space 16
--disable-log-requests
--dtype float16
--quantization fp8
--kv-cache-dtype fp8
--tensor-parallel-size {4 | 8}
--gpu-memory-utilization 0.9
--max-model-len 8192
--max-num-batched-tokens 8192
--num-scheduler-steps 1
--max-seq-len-to-capture 256
--max-num-seqs 256
--distributed-executor-backend mp
--enable-chunked-prefill
--enable-prefix-caching
```

Read the report at <https://facts.pt/UI0EJ07>

This project was commissioned by Dell Technologies.



Facts matter.®

Principled Technologies is a registered trademark of Principled Technologies, Inc. All other product names are the trademarks of their respective owners.

#### DISCLAIMER OF WARRANTIES; LIMITATION OF LIABILITY:

Principled Technologies, Inc. has made reasonable efforts to ensure the accuracy and validity of its testing, however, Principled Technologies, Inc. specifically disclaims any warranty, expressed or implied, relating to the test results and analysis, their accuracy, completeness or quality, including any implied warranty of fitness for any particular purpose. All persons or entities relying on the results of any testing do so at their own risk, and agree that Principled Technologies, Inc., its employees and its subcontractors shall have no liability whatsoever from any claim of loss or damage on account of any alleged error or defect in any testing procedure or result.

In no event shall Principled Technologies, Inc. be liable for indirect, special, incidental, or consequential damages in connection with its testing, even if advised of the possibility of such damages. In no event shall Principled Technologies, Inc.'s liability, including for direct damages, exceed the amounts paid in connection with Principled Technologies, Inc.'s testing. Customer's sole and exclusive remedies are as set forth herein.