Direct from Development

1S PowerEdge R7515 has Equivalent T4 GPU Performance to 2S PowerEdge R7425

Distinguished Next Gen AMD EPYC™ CPU

The launch of AMDs 2nd Generation EPYC™ (Rome) CPUs shook up the CPU industry by refining their proprietary Zen microarchitecture to new limits. With up to 64 cores, twice the amount of its predecessor (Naples), AMD went above and beyond the traditional tech mold by delivering a product truly worth of the term “next-gen”.

![AMD Rome CPU architecture graphic](image)

Figure 1 – AMD Rome CPU architecture graphic (large I/O die in the center with 8 chip dies containing 8 cores bordering the I/O die)

From a component-spec standpoint, the Rome CPU is 2x as capable as the Naples CPU. However, Dell Technologies wanted to confirm its ability to manage dense workloads that stress the processor. This led to various tests executed on the PowerEdge R7515 server, which supports 1 Rome CPU, and the PowerEdge R7425 server, which supports 2 Naples CPUs, to record and compare the performance of each CPU generation. Object detection, image classification and machine translation workloads were run with the support of NVIDIA T4 GPUs assisting the CPU(s).

VDI, IVA and Inference Studies

By executing tests on both servers (Figure 2) for various workloads (Figures 3-7), two factors are examined:

1. How the R7515 (Rome) and R7425 (Naples) solutions performed across various Machine Learning inference workloads. This accounts for the reduction of eight memory modules in the R7515 solution.
2. How NVIDIA T4 GPU performance compared between both solutions (QPS and inputs per second).
Server Details

<table>
<thead>
<tr>
<th>Division</th>
<th>system name</th>
<th>host processor model name</th>
<th>host processor core count</th>
<th>host memory capacity</th>
<th>host storage capacity</th>
<th>host processor interconnect</th>
<th>accelerators per node</th>
<th>accelerator model name</th>
<th>accelerator frequency</th>
<th>accelerator_memory_capacity</th>
<th>accelerator_memory_configuration</th>
<th>cooling</th>
<th>bus type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6xT4_Dell_IP12 4xT4_Dell_IP13/2</td>
<td>AMD EPYC 7561 32-Core Processor @201-Hz AMD EPYC 7762P 64-Core Processor @3.35Ghz</td>
<td>32</td>
<td>512 GB</td>
<td>2TB</td>
<td>NVMe SSD</td>
<td>4</td>
<td>NVIDIA Tesla T4</td>
<td>1500 MHz</td>
<td>16 GB</td>
<td>HBM2</td>
<td>Passive</td>
<td>ECC OFF</td>
</tr>
</tbody>
</table>

Figure 2 – Server configuration details for the 32-core server (left) and 64-core server (right)

Image Classification Performance

Figure 3 – Queries / Inputs per second for MobileNet-v1 model and ImageNet 224x224 dataset

Figure 4 – Queries / Inputs per second for ResNet-50 v1.5 model and ImageNet 224x224 dataset

Object Detention Performance

Figure 5 – Queries / Inputs per second for SSD w/ MobileNet-v1 model and COCO dataset

Figure 6 – Queries / Inputs per second for SSD w/ ResNet-34 model and COCO 1200x1200 dataset

Machine Translation

Figure 7 – Queries / Inputs per second for GNMT model and WMT E-G dataset
The figures above display the performance comparison of a 1S PowerEdge R7515 configured with 4 NVIDIA T4 GPUs and a 2S PowerEdge R7425 with 6 NVIDIA T4 GPUs. Although the bar graphs may not appear equivalent, once the total queries and inputs per second are divided by the total GPU count, we see that the performance per individual GPU is nearly equivalent (see Figure 8).

Now that the data is reduced to a common denominator of one GPU, the performance variance becomes easy to interpret. The inputs per second for Image Classification and Object Detection are nearly identical between server configurations; staying within ±3% of one another. Machine Translation numbers, however, are heavily boosted by the AMD Rome CPU. The queries per second (QPS) are a little more variant but are still very similar. All workloads stay within ±7% of one another, except for the object detection workload ResNet-34, which has a -14.53% loss in performance.

**Major Takeaways**

This data proves that despite executing the workload on a single socket server, the Rome server configuration is still executing vision and language processing tasks at a nearly equivalent performance to the Naples configuration. Knowing this, Dell Technologies customers can now be informed of the following takeaways upon their next PowerEdge configuration order:

1. A single socket 64-core AMD Rome CPU performs at near equivalence to two socket 32-core AMD Naples CPUs for vision and language processing tasks. This means that inference workloads in the AI space will be able to perform effectively with less components loaded in the server. Therefore, customers running workloads such as inference that are not impacted by a reduction in total system memory capacity would be great candidates for switching from 2S to 1S platforms.
2. Non-Uniform Memory access (NUMA) memory and I/O performance issues associated with 2S platforms is avoided with the 1S R7515 Rome configuration. This is beneficial to I/O and memory intensive workloads as data transfers are localized to one socket; therefore avoiding any associated latency and bandwidth penalties.
3. 64-core single socket servers typically offer better value due to the amortization of system components.
4. Reducing the number of CPUs and memory will reduce the total power consumption.
Conclusion

One 2nd Generation AMD EPYC™ (Rome) CPU is capable of supporting AI vision and language processing tasks at near-equivalent performance to two 1st Generation AMD EPYC™ (Naples) CPUs. The advantages attached to this generational performance gap, such as increased cost-effectiveness, will appeal to many PowerEdge users and should be considered for future solutions.