|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Change Request** | | | | | | | |
| **Document** | **ORAN-WG6.CLOUD-REF-B** | **ver** | **01.00.00** | **CR** | **NVD-001** | **rev** | **4** | |

|  |  |  |  |
| --- | --- | --- | --- |
| ***Title:*** | GPU based Edge Cloud Server Reference Configuration | | |
| ***Source to WG:*** | NVIDIA | | |
| ***Target WG :*** | **WG6** | | |
| ***Category:*** | **B** | ***CR Creation Date*** | July 10, 2020 |
|  | *Use one of the following* ***categories****:* ***A*** *(mirror corresponding to a change in an earlier release)* ***B*** *(addition of feature),* ***C*** *(functional modification of feature)* ***D*** *(editorial modification)* ***F*** *(correction)*  Detailed explanations of the above categories can be found in 3GPP [TR 21.900](http://www.3gpp.org/ftp/Specs/html-info/21900.htm). | | |

|  |  |
| --- | --- |
| ***Reason for Change:*** | To include reference design for edge cloud server with GPU based acceleration |
| ***Summary of change:*** | New text and table are proposed and can be reviewed by track change in the text below |
| ***Consequences if not aproved:*** | Spec. will lack diversity in edge cloud server reference design catering different HW accelerators and accelration profiles |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Clauses affected:*** | <list specific document sections impacted by the CR> | | | | |
|  | **Y** | **N** |  | |  |
| ***Other specs*** |  | **X** | Other core specifications: | <fill in related CRs if “Y”> | |
| ***affected:*** |  | **x** | Test specifications: | <fill in related CRs if “Y”> | |
| ***(show related CRs)*** |  | **X** | O&M Specifications: | <fill in related CRs if “Y”> | |
| ***Supporting material:***  ***Other comments:*** | <provide file name or URL of any material supporting this CR> | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| ***Status:*** |  | ***CR Closed Date:*** |  |
| ***Outcome:*** |  | ***Duplication:*** |  |
| ***Outcome explanation:*** |  | | |

The proposed changes are indicated by Track Changes in the text below.

# Cloud Platform Reference Design Example

## Cloud Platform HW Reference Design

This section specifies the cloud platform hardware with a focus on server and accelerator hardware. These reference

designs are examples, not normative. This document focuses on x86 processor-based cloud server design; it will address ARM processor-based designs in a future release.

### Regional Cloud Server

### Edge Cloud Server

[…]

Table 5-3 shows an example reference design for the edge cloud servers with GPU based acceleration.

|  |  |
| --- | --- |
| Form factors | Depth ≤ 450mm, height ≤ 2RU, width ≤ 19in |
| Processor | 12C x86 CPU complex or better (e.g. Xeon) |
| Memory | 64 GB, 2400 MHz; or higher |
| Expansion Slots | At least 1x HHHL and 1x FHFL Dual Slot |
| Drive Bays | 6x hot-swappable 2.5-inch SATA or SAS drives; 2x NVMe drives; all drives accessible from front panel |
| RAID support | RAID1 |
| Network Interface | 2x 100/50/25/10 GbE ports, 1 dedicated 1GbE management port, 1x [Mellanox ConnectX](https://www.mellanox.com/sites/default/files/doc-2020/pb-connectx-6-dx-en-ic.pdf)[®](https://www.mellanox.com/sites/default/files/doc-2020/pb-connectx-6-dx-en-ic.pdf)[-6 Dx](https://www.mellanox.com/sites/default/files/doc-2020/pb-connectx-6-dx-en-ic.pdf) (with crypto offload support, eCPRI offload with 5T for 5G technology)  Efficient peer-to-peer support, either through root complex, or external PCIe switch (such as ExpressLane™ PEX 8747) recommended for high bandwidth use cases (like multi-carrier 100MHz massive MIMO systems) |
| PSU | * 1+1 redundant * Hot-swappable * Support for 220V AC or -48V DC * PSUs are connected through the rear panel (AT&T, China Mobile) or front panel (China Telecom, DT, KDDI, Orange) |
| Temperature | -5 °C ~ 55 °C (-10 °C to 60 °C preferred) |
| Security | TPM |
| Availability | Hot-swap fans/PSUs; 45℃ continuous operation; light path diagnostic LEDs; front-access diagnostics via dedicated USB port |
| Manageability | Support for IPMI, SNMP, and Redfish |
| Acceleration  (inline, end-to-end HIGH PHY profile) | [NVIDIA](https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/tesla-t4/t4-tensor-core-product-brief.pdf)[®](https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/tesla-t4/t4-tensor-core-product-brief.pdf) [T4](https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/tesla-t4/t4-tensor-core-product-brief.pdf), PCIe x16/x8 Gen3, HHHL (for smaller capacity, low band deployments).  [NVIDIA](https://images.nvidia.com/content/tesla/pdf/Tesla-V100-PCIe-Product-Brief.pdf)[®](https://images.nvidia.com/content/tesla/pdf/Tesla-V100-PCIe-Product-Brief.pdf) [Tesla](https://images.nvidia.com/content/tesla/pdf/Tesla-V100-PCIe-Product-Brief.pdf)[®](https://images.nvidia.com/content/tesla/pdf/Tesla-V100-PCIe-Product-Brief.pdf) [V100](https://images.nvidia.com/content/tesla/pdf/Tesla-V100-PCIe-Product-Brief.pdf), PCIe x16 Gen3, FHFL (for higher capacity, to enable support for multi-sector massive MIMO (100 MHz and above) with advanced beam-forming techniques). |

Table 5‑3. Edge cloud server example reference design with GPU acceleration.