# SDP Hardware Decomposition View

Contributors: C. Broekema

## TABLE OF CONTENTS



## LIST OF ABBREVIATIONS

| | |
|---|---|
| **AA** | Aperture array |
| **BGP** | Border Gateway Protocol |
| **BMC** | Baseboard Management Controller |
| **COTS** | Common Off The Shelf |
| **CSP** | Central Signal Processor |
| **DC** | Direct Current |
| **FPGA** | Field Programmable Gate Array |

| | |
|---|---|
| **GbE** | Gigabit Ethernet |
| **GPGPU** | General Purpose Graphics Processor Unit |
| **HCA** | Host Channel Adaptor |
| **ICD** | Interface Control Document |
| **IP** | Internet Protocol |
| **MAID** | Massive Array of Idle Drives |
| **NIC** | Network Interface Controller |
| **NSDN** | Non-Science Data Network |
| **PBS** | Product Breakdown Structure |
| **PDU** | Power Distribution Unit |
| **RAM** | Reliability Availability Maintainability (analysis) |
| **SDP** | Science Data Processor |
| **TCP/IP** | Transmission Control Protocol / Internet Protocol |
| **TM** | Telescope Manager |
| **UDP/IP** | User Datagram Protocol / Internet Protocol |
| **VLBI** | Very Long Baseline Interferometry |

## 1. Primary Representation

The SDP hardware decomposition is represented by the indented list below. This shows the complete SDP hardware PBS. Figure 1 shows how the hardware is physically organised. Note the hierarchical nature clearly shown in Table 1. We do not make any distinction between the MID and LOW hardware. While we expect the MID and LOW SDP hardware implementations to differ in details such as size and number of specific components, the high level architecture and decomposition on smaller components will be identical.

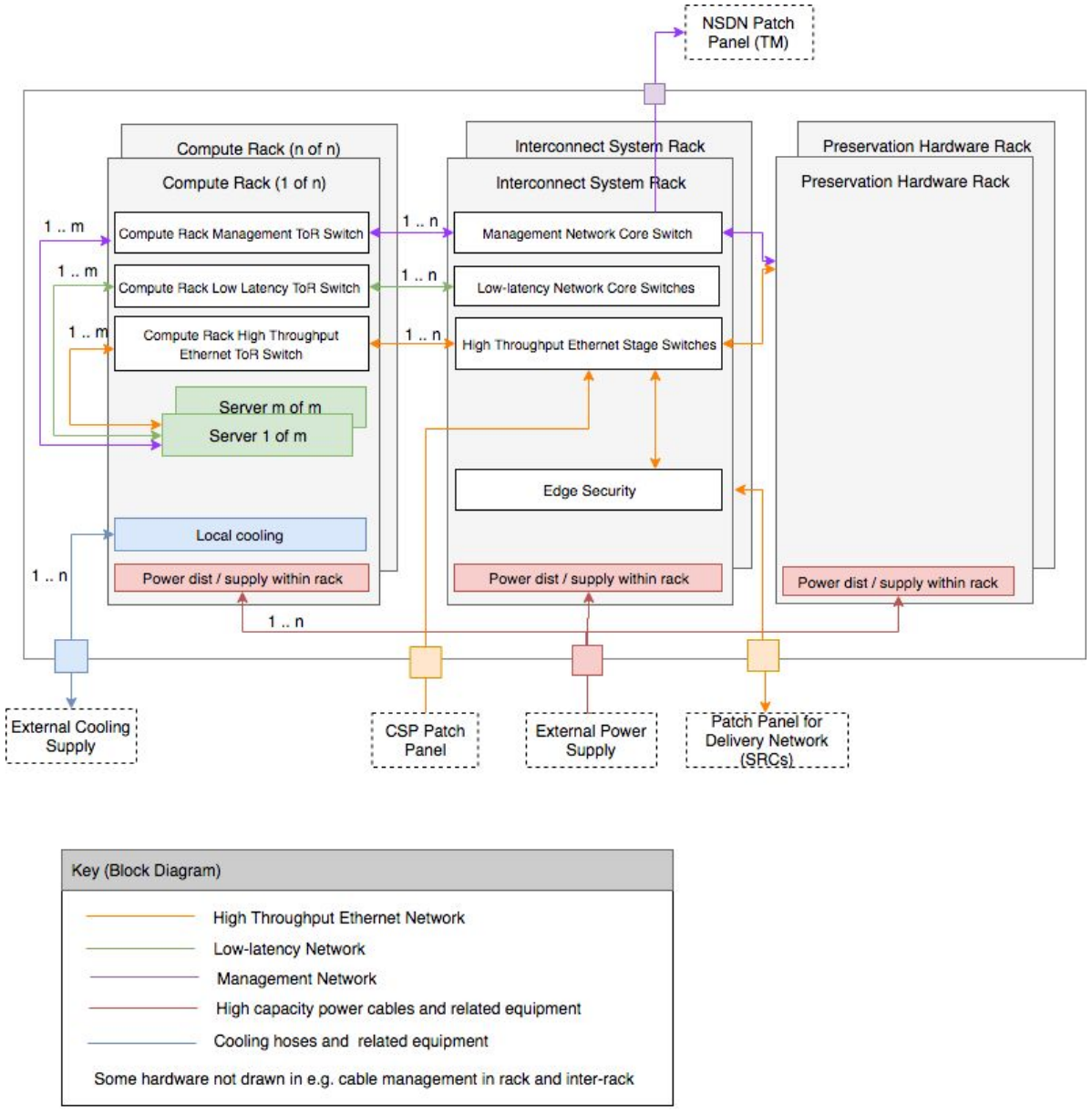


**Figure 1:** SDP hardware block diagram

This hierarchical nature of the hardware is key to the scalability of the SDP. We see the SDP Compute Rack as a self-contained resource pool that can be procured as a unit. There is no practical limit to the number of Compute Racks that can be installed, apart from increased cost in the interconnect system and a general limit in available energy and funds. Compute racks are heterogeneous, both between racks and potentially within a compute rack. This is explored in more detail in Section 5.3.

304-000000 Science Data Processor (SDP) MID

- 304-000001 SDP Compute Hardware MID
  - 304-000003 Compute Rack
    - P.2.1.1.1 Compute Rack Management Switch
    - P.2.1.1.2 Server
      - P.2.1.1.2.1 Latency optimized cores
      - P.2.1.1.2.2 Main memory
      - P.2.1.1.2.3 Throughput optimized cores
      - P.2.1.1.2.4 High Throughput Ethernet NIC
      - P.2.1.1.2.5 Low Latency network NIC
      - P.2.1.1.2.6 Out-of-band management unit (BMC)
      - P.2.1.1.2.7 Storage
        - P.2.1.1.2.7.1 Capacity Storage
        - P.2.1.1.2.7.2 Performance Storage
      - P.2.1.1.2.8 Management Network NIC
    - P.2.1.1.3 Compute Rack High Throughput Ethernet Switch
    - P.2.1.1.4 Compute Rack Low Latency Switch
    - P.2.1.1.5 Cabling
    - P.2.1.1.6 Racks Infrastructure
      - P.2.1.1.6.1 PDU
      - P.2.1.1.6.2 Bulk power supply (possibly, may cross racks)
      - P.2.1.1.6.3 Cable management
      - P.2.1.1.6.4 Local cooling (i.e. water cooled rack doors)
      - P.2.1.1.6.5 Rack
  - 304-000004 Interconnect System MID
    - P.2.1.2.1 Low Latency Network Core Switch
    - P.2.1.2.2 Management Network Core Switch
    - P.2.1.2.3 High Throughput Ethernet Core Switch
      - P.2.1.2.3.1 Pluggable optics
      - P.2.1.2.3.2 Fibre
      - P.2.1.2.3.3 Copper cabling
      - P.2.1.2.3.4 Patch panel
    - P.2.1.2.4 Edge Security
    - P.2.1.2.5 Interconnect System Rack
  - 304-000005 Inter-rack Infrastructure MID
    - P.2.1.3.1 Hot Aisle / Cold Aisle equipment (roofs, doors)
    - P.2.1.3.2 High capacity power equipment (cables, etc)
    - P.2.1.3.3 Liquid cooling equipment (hoses, etc)
    - P.2.1.3.4 Cross-rack cable management (trays, etc)
- 304-000002 SDP Preservation Hardware
    - 304-000006 Hierarchical Storage Management
    - 304-000007 Intermediate Storage
    - 304-000008 Long Term Storage

104-000000 Science Data Processor (SDP) LOW

- 104-000001 SDP Compute Hardware LOW
  - 104-000003 Compute Rack LOW
    - P.2.1.1.1 Compute Rack Management Switch
    - P.2.1.1.2 Server
      - P.2.1.1.2.1 Latency optimized cores
      - P.2.1.1.2.2 Main memory
      - P.2.1.1.2.3 Throughput optimized cores

- P.2.1.1.2.4 High Throughput Ethernet NIC
- P.2.1.1.2.5 Low Latency network NIC
- P.2.1.1.2.6 Out-of-band management unit (BMC)
- P.2.1.1.2.7 Storage
  - P.2.1.1.2.7.1 Capacity Storage
  - P.2.1.1.2.7.2 Performance Storage
- P.2.1.1.2.8 Management Network NIC

- P.2.1.1.3 Compute Rack High Throughput Ethernet Switch
- P.2.1.1.4 Compute Rack Low Latency Switch
- P.2.1.1.5 Cabling
- P.2.1.1.6 Racks Infrastructure
  - P.2.1.1.6.1 PDU
  - P.2.1.1.6.2 Bulk power supply (possibly, may cross racks)
  - P.2.1.1.6.3 Cable management
  - P.2.1.1.6.4 Local cooling (i.e. water cooled rack doors)
  - P.2.1.1.6.5 Rack

- 104-000004 Interconnect System MID
  - P.2.1.2.1 Low Latency Network Core Switch
  - P.2.1.2.2 Management Network Core Switch
  - P.2.1.2.3 High Throughput Ethernet Core Switch
    - P.2.1.2.3.1 Pluggable optics
    - P.2.1.2.3.2 Fibre
    - P.2.1.2.3.3 Copper cabling
    - P.2.1.2.3.4 Patch panel
  - P.2.1.2.4 Edge Security
  - P.2.1.2.5 Interconnect System Rack
- 104-000005 Inter-rack Infrastructure MID
  - P.2.1.3.1 Hot Aisle / Cold Aisle equipment (roofs, doors)
  - P.2.1.3.2 High capacity power equipment (cables, etc)
  - P.2.1.3.3 Liquid cooling equipment (hoses, etc)
  - P.2.1.3.4 Cross-rack cable management (trays, etc)

- 104-000002 SDP Preservation Hardware
  - 104-000006 Hierarchical Storage Management
  - 104-000007 Intermediate Storage
  - 104-000008 Long Term Storage

**Table 1:** The SDP hardware product breakdown structure. This shows the complete hierarchical hardware breakdown of the SDP. Note that not all of these products are required or expected to be implemented for all elements.

Figure 2 shows a logical view of the primary physical SDP networks: high throughput Ethernet networks, and the management network. Not shown here is the low latency interconnect network that provides a low latency and high bandwidth network between all Servers, but does not extend beyond the Servers.

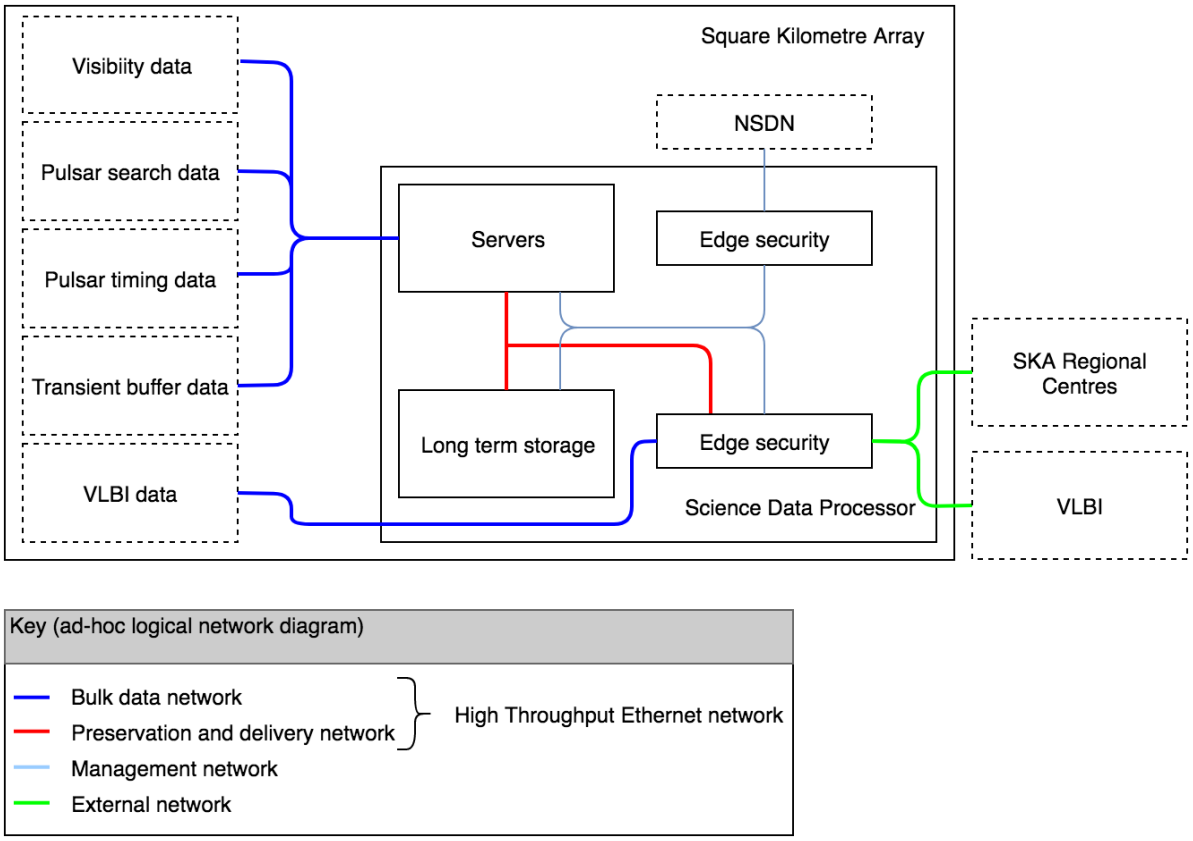


**Figure 2:** A logical view of the SDP networks

Although the high throughput Ethernet network is a single product group in our PBS, we have drawn it as two distinct networks in Figure 2. Due to performance and and associated interference concerns, we may implement these two networks physically separate.

The Bulk data network is used to transport data from the instrument (correlator, pulsar search and pulsar timing machines at CSP and transient buffer data from the receivers) to the SDP. This uses 100GbE over the long haul lines to the SDP switches, and 25GbE to the Servers. The vast majority of the data consists of UDP/IP streams (visibility data and possibly transient buffer data), with the rest TCP/IP. We refer to the relevant ICDs for exact data volumes and structures [AD11, AD07].

The management network is a generic Ethernet network used for non-application data, such as monitoring and control data. This network is also responsible for the distribution of application code and images to all servers.

We show the external network to the SKA Regional Centres for completeness, however this is formally out of scope for the SDP. We currently assume this to be based on long haul 100GbE technology.

## 2. Element Catalogue

### 2.1. Elements and Their Properties

The Science Data Processor hardware is built up in an hierarchical fashion, as shown in Figure 1. In this section we enumerate the various elements that make up the Science Data Processor. For clarity we describe the Server elements separately in Section 2.1.2.

### 2.1.1 P.2.1 SDP Compute Hardware (104-000001, 304-000001)

- P.2.1.1 Compute Rack (104-000003, 304-000003)
  A Compute Rack is the basic replicable unit in the SKA SDP. It is a self-contained, independent collection of servers and infrastructure needed to keep the compute rack running.
  The compute rack is expected to be a monolithic procurable unit. Considering the extended timeframe of both construction and operations, it is expected that compute racks from different procurement cycles are heterogeneous. Within a procurement, compute racks may be identical to ease operational load and scheduling complexity. Compute racks may differ between instruments, for instance in the number and characteristics of the containing servers.
- P.2.1.1.1 Compute Rack Management Switch
  This is/are the switch(es) that connect(s) the core management network to the components in the compute rack. This network encompasses both in-band as well as out-of-band management. Data volumes on this network are still subject of discussions, but it seems likely that at least the in-band component will require more capable networks (i.e. >1GbE).
- P.2.1.1.2 Server
  The basic building block of the SDP Compute Rack, described in more detail below.
- P.2.1.1.3 Compute Rack High Throughput Ethernet Switch
  An Ethernet (top-of-rack) switch that connects the high throughput ethernet network core switch to the server high throughput ethernet NIC. This is required to be Ethernet. The streaming and non-reliable nature of the data from CSP means that special care needs to be taken in the selection of this component to ensure data loss is minimised. Software-defined networking capability is currently not a requirement, but is investigated for viability. IP Multicast capability is currently not a requirement, but is investigated for viability. BGP capability is currently not a requirement, but is investigated for viability.
- P.2.1.1.4 Compute Rack Low Latency Switch
  The compute rack level low-latency interconnect switch connects the core low-latency interconnect network to the compute server based low-latency interconnect HCA.
- P.2.1.1.5 Cabling
  This is the collection of cables and ancillary equipment needed to connect all compute rack components. This includes both power and network cables (or fibres, as needed), as well as optics where necessary.
- P.2.1.1.6 Racks Infrastructure
  Per rack infrastructure products.
- P.2.1.1.6.1 PDU
  Rack-based power distribution units. Interface with the bulk power delivery from the local infrastructure.
- P.2.1.1.6.2 Bulk power supply (possibly, may cross racks)
  This optional product may be used to distribute DC power to servers. This is preparation of a possible move to rack level power supplies instead of server based power supplies. The design of this is unclear, since there currently is no current industry standard implementation.
- P.2.1.1.6.3 Cable management
  Rack-based equipment to manage cables in a compute rack.
- P.2.1.1.6.4 Local cooling (i.e. water cooled rack doors)
  These are rack level thermal solutions that supplement the data center level cooling where necessary. This may include water/fluid cooled rack doors. This may require a (bi-directional) fluid interface with the infrastructure.
- P.2.1.1.6.5 Rack

These are the racks that contain the components mentioned above. These are likely industry standard 19 inch racks, although a different standard may be adopted if this becomes useful. Note that a integrated solution is still an option, which would make this component rather specialised.
Rack level power distribution is included. We expect servers to require normal mains voltage power, although developments of efficient lower voltage DC power distribution are followed with interest.

- P.2.1.2 Interconnect System (104-000004, 304-000004)
- P.2.1.2.1 Low Latency Network Core Switch
  Centralised core switch(es) for the Low Latency network
- P.2.1.2.2 Management Network Core Switch
  Centralised core switch(es) for the management network
- P.2.1.2.3 High Throughput Ethernet Core Switch
  Centralised core switch(es) for the high throughput ethernet network
- P.2.1.2.3.1 Pluggable optics
  Optical transceivers for our various networks, where necessary. Our working assumption is that we use electrical communication wherever possible, for cost reasons. This assumption may need to be reviewed at some later stage.
- P.2.1.2.3.2 Fibre
  Fibre optic cables for our various networks. Our working assumption is that we use electrical communication over copper cables wherever possible, for cost reasons. This assumption may need to be reviewed at some later stage.
- P.2.1.2.3.3 Copper cabling
  Copper cabling, including power delivery and networking.
- P.2.1.2.3.4 Patch panel
  Patch panels for inter-connecting the various compute racks and networking infrastructure.
- P.2.1.2.4 Edge Security
  Device(s) and/or other security measures handling access control and security for connections from outside of SDP, including non-science data network and SKA Regional Centres.
- P.2.1.2.5 Interconnect System Rack
- P.2.1.3 Inter-rack Infrastructure (104-00005, 304-000005)
  Hardware that crosses a single rack.
- P.2.1.3.1 Hot Aisle / Cold Aisle equipment (roofs, doors)
  Equipment to support proper thermal household of the data centre, if not covered by the hosting institute. This may include automatic doors and rack roofs for hot and cold aisles, as well as rack blanks for airflow management.
- P.2.1.3.2 High capacity power equipment (cables, etc)
  Energy delivery equipment to connect the local infrastructure energy delivery to the rack-based power-distribution units.
- P.2.1.3.3 Liquid cooling equipment (hoses, etc)
  Equipment to connect liquid cooled rack-based equipment, such as liquid cooled rack doors, to the hosting
- P.2.1.3.4 Cross-rack cable management (trays, etc)
  Inter-rack equipment, such as roof-mounted trays, to manage cables and fibres between compute racks.
- P.2.2 SDP Preservation Hardware (104-000002, 304-000002)
  The Preservation Hardware is anticipated to be a COTS platform sourced separately. The Hierarchical Storage Management will provide an interface to Cold Buffer and manage the intermediate and long term storage of the Preservation Hardware which could consist of

MAID and Tape. The longevity of data products on these media will be dictated by access and policy rules.

- P.2.2.1 Hierarchical Storage Management (104-000006, 304-000006)
  Software to handle automatic migration from intermediate to long term storage and vice versa. Also provides the interface to the various other components.
- P.2.2.2 Intermediate Storage (104-000007, 304-000007)
  This is the performant tier of the hierarchical storage stack in the SDP preservation hardware.
- P.2.2.3 Long Term Storage (104-000008, 304-000008)
  This is the capacity tier of the hierarchical storage stack in the SDP preservation hardware. This is intended to be low-cost, high-capacity storage that can cheaply and efficiently store SDP products for the lifetime of the telescope.

### 2.1.2 P.2.1.1.2 Server

The server is the smallest decomposable unit in the SKA Science Data Processor. A server consists of a number of components, described below, not all of which are required for all servers. In Section 4 we describe how different servers may require different server components.

- P.2.1.1.2.1 Latency optimized cores
  These are cores optimised for latency sensitive operations, such as receiving large volumes of streaming data and real-time processing of that data. Generally these are large superscalar, out-of-order cores with multiple stages of cache, found in mainstream products from Intel, IBM, AMD and various ARM licensees.
- P.2.1.1.2.2 Main memory
  High performance, usually volatile and relatively small capacity, server storage used to temporarily store program state and data. Compare to P.2.1.1.2.7, which is lower performance, higher capacity and usually non-volatile server storage used to store in. This includes all memory within a server, as well as any memory on included accelerators, if applicable. It is likely that the boundary between this product and P.2.1.1.2.7 will blur in the next decade or so, which will mean hardware may be shared between these two products. The different applications, i.e. program memory versus (temporary) data storage, will mean these products likely will remain to have their own place in servers.
- P.2.1.1.2.3 Throughput optimized cores
  These are computational cores optimised for throughput rather than latency. It is expected that the bulk of our required computational resources will be delivered by such cores. They are typically massively parallel accelerator type cores. Multiple levels of internal parallelism and some form of Single Instruction Multiple Data (SIMD) are common in these kinds of devices. While we make no assumptions on the sort of device in this architecture, current examples are GPGPU accelerators, like those available from Nvidia and AMD. We note that FPGA based accelerators may also deliver suitable throughput optimised cores.
- P.2.1.1.2.4 High Throughput Ethernet NIC
  This Network Interface Controller is connected to the Bulk data network and handles ingress of data into the SDP. The ingress data stream is largely based on UDP/IP over Ethernet, with a small additional portion being TCP/IP over Ethernet. This will be a Ethernet NIC. The Ingress NIC may potentially share hardware with the egress NIC.
  This NIC is also connected to the preservation and delivery network, which handles data egress from the processor platform into preservation. This may share hardware with the ingress NIC, as described above, since this is also a Ethernet network. Traffic is, like ingress, mostly unidirectional.

Note that for performance reasons this single product may be implemented more than once in a single Server.

- P.2.1.1.2.5 Low Latency network NIC
  The low-latency network is connected to the low-latency interconnect HCA. While this name may imply a discrete adapter added to the server, we make no choice at this stage that this must be the case. Indeed, further integration of components into the CPU seems likely, which may lead to these HCAs being integrated into the CPU and the associated connections being landed on board.
- P.2.1.1.2.6 Out-of-band management unit (BMC)
- P.2.1.1.2.7 Storage
  Storage capacity in the server, used both for program and operating system storage, as well as for buffering instrument data. Can be divided into capacity and performance storage, based on performance, capacity and cost.
- P.2.1.1.2.7.1 Capacity Storage
  This is capacity optimised storage. Compare to the performance storage, which is performance optimised. This is where we find the cold buffer.
- P.2.1.1.2.7.2 Performance Storage
  This is performance optimised storage. Compare to the capacity storage, which is capacity optimised. This is what the hot-buffer is allocated on.
- P.2.1.1.2.8 Management Network NIC

## 2.3. Element Interfaces

The physical external interfaces on the hardware are described in the Interface control documents with SaDT [AD11, AD07].

# 3. Context Diagram

See Figure 1 and Figure 2.

## 4. Variability Guide

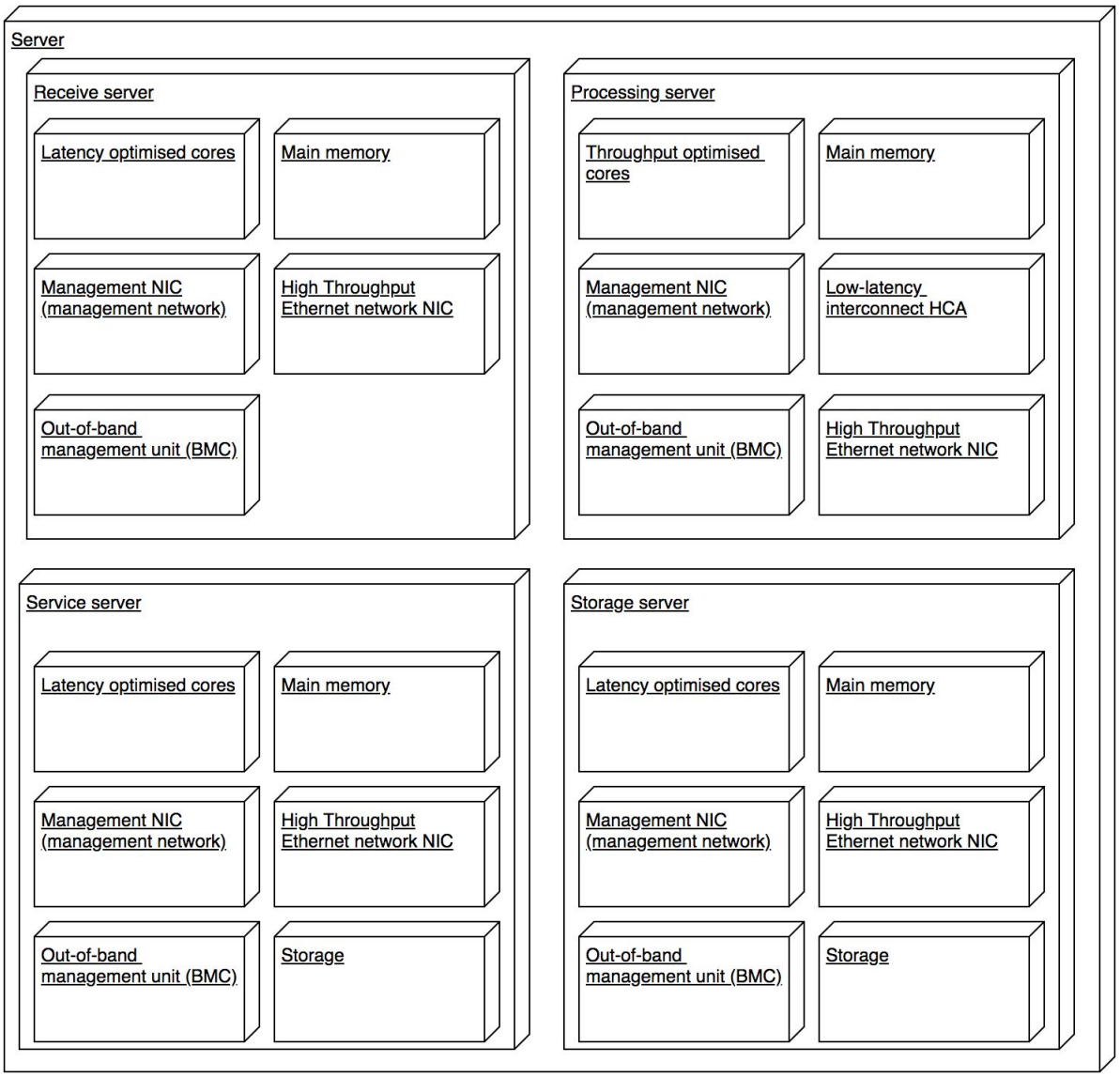


**Figure 3:** Server decomposition

A critical component in the SDP Compute Rack, shown in Figure 1, is the server. In this figure we show how the server can be de composed into four *personalities,* described in more detail below (section 5.3). Each of these consists of a number of required components, shown here.

Figure 4 shows a non-exhaustive list of possible Server variability. Apart from the server personalities discussed in section 5.3 below, physical heterogeneity in Servers will inevitably occur, both due to staged roll-out and the desire to optimise part of the Server inventory for particular applications or capabilities.

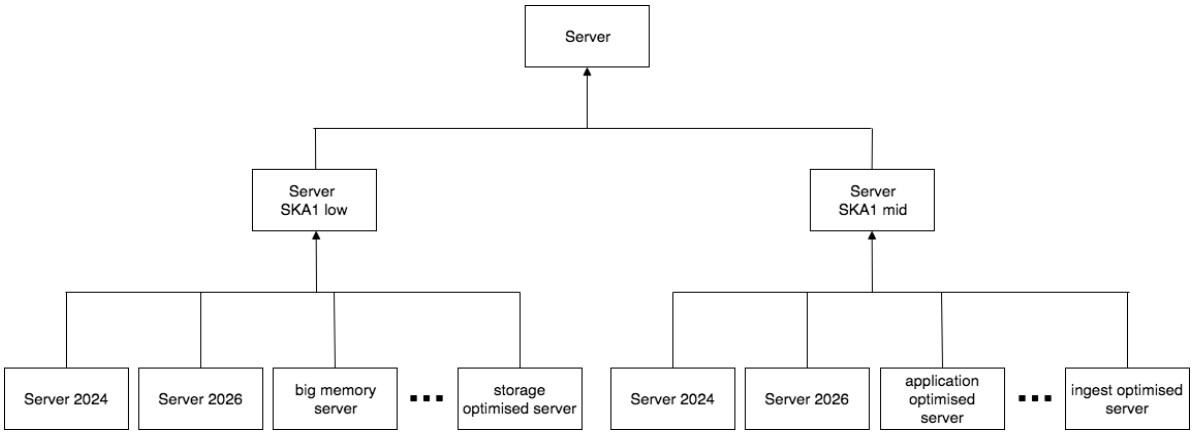


**Figure 4**: SDP Server generalisation view.

This figure shows that servers may differ in characteristics between instruments, as well as within an instrument. This heterogeneity may be due to advancing technologies (and inherently time), due to instrument specific optimisation, or due to application specific tuning. An non-exhaustive list of possible variations is shown in this figure. Note that for operational reasons we will strive for a high degree of homogeneity wherever possible and viable.

This will result in a heterogeneous Server ecosystem, with, among others:

- Variability over time / procurement cycles
- Variability between instruments
- Variability within a procurement cycle (i.e. specialised servers, specialised compute racks, etc)

## 5. Rationale

In this section we describe some of the reasoning behind the choices made for SDP hardware.

### 5.1 Primary rationale

The SKA SDP requires a pool of operational compute, storage and data transfer capacity in order to receive instrument data from the SKA instrument, and process it into science ready data products to be exported to the SKA Regional Centres. To accommodate this, SDP will consist of a number of compute racks, each with a number of servers (as described above) that deliver the required resources. The hierarchical nature of the SDP hardware is shown in this hardware decomposition view, but not necessarily exposed to the software layers above this. For operational and procurement reasons, such hierarchical structuring is convenient. We can imagine for instance that a staged roll-out will be done at compute rack level, with each procurement adding or replacing a number specific number of compute racks. We've tried to show this variability in Figure 4.

We note that the hardware description is deliberately very vague. This has a number of reasons, chief of which is the long time until procurement. Six to seven years from today, the hardware landscape may be markedly different from what can be bought today. We do not think describing a hardware design in detail with today's technology is worthwhile, although we have costed based on current day technology extrapolated to the roll-out phase [RD01].

### 5.2 SDP operational availability

An analysis of the operational availability of SDP hardware is discussed elsewhere [RD02]. We note however that the parallel nature of the SDP hardware, where we duplicate the same resources many times over, make the SDP inherently highly redundant, and therefore resilient against failure. In

other words, many components in SDP may (and indeed are expected to) fail, without bringing SDP as a whole down. We refer to the RAM analysis [RD02] for the definition of available and inherent availability.

## 5.3 Servers and personalities

The SKA SDP has four primary high level tasks

1. Receive and condition instrument data, including real-time processing
2. Batch processing of conditioned data
3. Buffer data
4. System management tasks

For efficiency reasons we identify four server personalities to match these high level tasks, as shown in Figure 4, each with a required subset of the hardware available to servers. These are:

- Receive -- receive streaming data from the instrument
- Processing -- process data in the buffer into SDP products
- Service -- handle routine tasks, such as monitoring and control, etc.
- Storage -- buffer data

Physical Servers have the capability to assume one or more of these personalities, based on their hardware configuration. One or more of these personalities are assigned to servers based on need and available hardware (see Section 4 and Figure 4 for a discussion on the variability of the hardware in SDP) and other resources, such as available energy or cooling capacity. Hardware variability defines the personalities that equipment can take on, for example: a storage optimised server can be a storage server, but may be less suitable for processing.

For efficiency reasons it is possible that a server assumes more than one personality at any one time, for instance a server may have throughput optimized and latency optimised cores available, and therefore assume both a receive and processing personality. We note that sharing personalities requires careful attention to avoid bottlenecking.

While different personalities require quite different hardware configurations within a Server, it is not inconceivable that for cost reasons all Servers within a compute rack will have identical hardware configurations. This may lead to idle hardware in Servers that have, for instance, a service personality. The optimal combination of different hardware configurations is a complex trade-off between cost, energy consumed by potentially idle hardware, and operational complexity that comes with having to support different types of nodes versus a homogeneous system.

## 5.4 Storage classes within the server

One of the key characteristics of the SDP compute profile is the split between real-time processing and batch processing, with a storage component, the buffer, between the two. We expect that for cost reasons this buffer will consist of a mix of high performance, but expensive *performance storage*, and lower performance and cheaper *capacity storage*. The ratio between the two, or indeed the need for the mix itself, will depend greatly on price and performance development over the coming years.

We note that mixing classes of storage makes the system more complex. If cost of a single class of storage is cost- and performance-wise feasible, this is the preferred option. Storage class memory, expected to become mainstream by the time we procure hardware for SDP, may shift this landscape sufficiently to allow a single class of storage to be affordable.

## 5.5 Networks

The high throughput Ethernet network shown in Figures 2 and 3 have to handle very different traffic profiles. Data from the instrument, i.e. visibilities from CSP, are sensitive to loss, due to the use of the unreliable UDP/IP protocol.

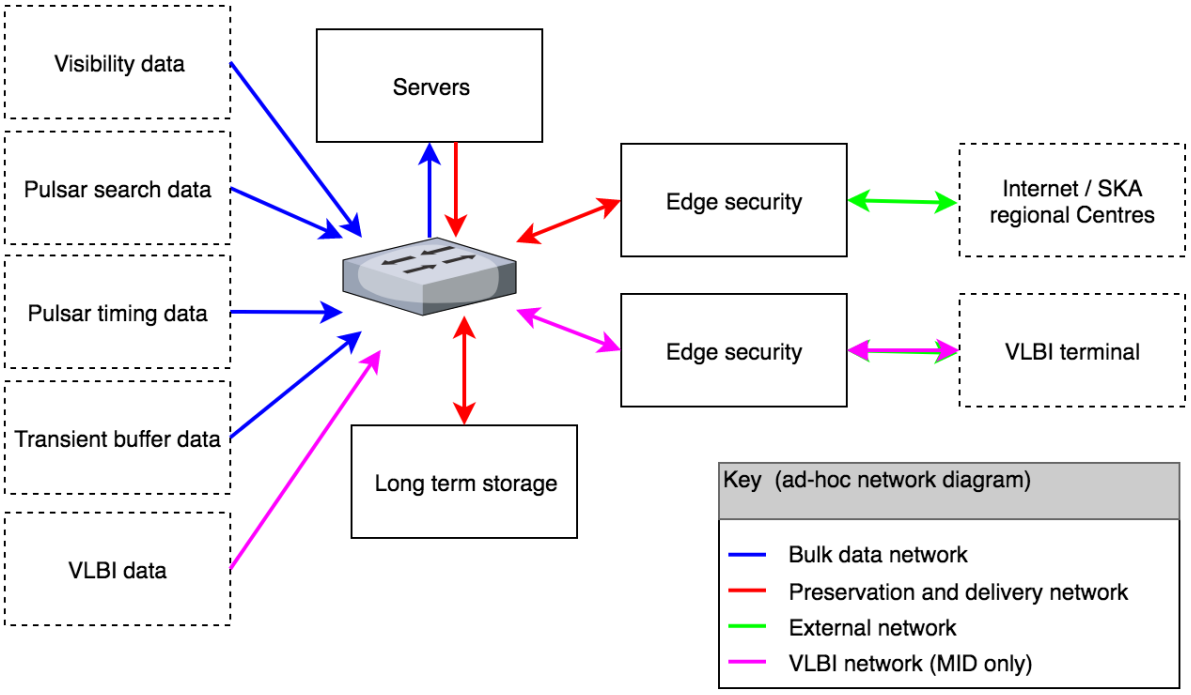


**Figure 5:** The SDP high throughput Ethernet network

Figure 5 shows the SDP high throughput Ethernet network. In this figure we show that, while this entire network may share a single technology, and indeed a lot of hardware, for performance reasons it is likely that these will be physically split into up to five separate networks. In this figure we identify

- Bulk data network -- instrument data from CSP and AA-low stations to servers
- Preservation and delivery network -- connecting servers to storage and edge security
- External network -- shown, but formally out of scope for SDP
- VLBI network -- VLBI data only transits the SDP network to a VLBI terminal, which is out of scope for SDP

For security, any device not fully under SDP control will only be connected through edge security.

The bulk data network and the preservation and delivery network in particular are likely to be physically kept separated. Both are high capacity, high throughput networks, that are sensitive to bottlenecks. Furthermore, on the bulk data network, visibility data (which accounts for most of the data transported) is transported using the unreliable UDP/IP protocol: which means that any lost data is irretrievable. We also note that the radically different communication patterns on these two networks may make it difficult to optimise switch and port configurations if they need to handle these two networks.

However, if experience shows that these two networks can peacefully coexist on the same hardware, this is a viable and cost-effective alternative.

### Switched bulk data network rationale

In Figure 5 we show that there is no direct connection between the SDP servers on the one hand, and the instrument data generating equipment at CSP (such as the correlator / beamformer and the

pulsar search and timing clusters) on the other. All traffic is directed through the high throughput Ethernet network, and in particular the bulk data network component of this.

This architecture, as opposed to directly connecting receive capable servers to the incoming network connections from CSP, was chosen for several reasons:

1. Flexibility: by having a switched bulk data network, we can freely choose which receive capable servers receive data from a CSP node. There is no need for a one-on-one relation between a CSP node and a physical SDP receive node.
2. Maintainability: If a receive capable server in SDP fails or is replaced at the end of its service life, the data streams from CSP can be received by another server without having to physically reconnect the network to a different server.
3. Sharing of links: the physical links between CSP and SDP are potentially shared between the correlator / beamformer, the pulsar timing and search systems and the VLBI data. In a switched bulk data network, these are naturally directed to their respective destinations.
4. Emerging technologies: programmable networks are an emerging technology that we have been investigating for some time.

The data transmission pattern in the bulk data network is fairly static, with continuously high loading of all links and large frames to maximise performance. This is very different from conventional network traffic, which switches are designed for. There is therefore a risk that the memory buffers in such switches are temporarily overloaded, which may lead to some data loss, as documented in SDPRISK-324. This risk is well understood and can usually be mitigated with careful tuning of individual buffer tuning. We note that switches with extra large buffers are available in the market today that specifically address this risk.

### 5.6 Memory types

Memory in SDP hardware may consist of various types and technologies, depending on the implementation and hardware developments. Current state of the art host + accelerator servers would distinguish between

- Host memory, including main memory and several levels of caching
- Device memory, divided into several classes depending on accelerator type and vendor

In future servers, which may be procured for SDP, these may be combined, or more types may be added. An example of the first are servers where latency and throughput optimised cores are combined in the same package, sharing memory, such as Nvidia Tegra solutions or AMD APUs (the AMD exascale concept is based on this concept). In addition to the types above, we may add storage class memory to the host to act as a fast storage medium, or an abundant, non-volatile but relatively slow memory path.

In general we expect the current differentiation between host and device memory to disappear. Instead we expect to see a number of stages, similar to today's stack of caches and main memory, with added elements such as storage class memory and remote but transparently accessible memory on an accelerator.

## 6. Related views

*SDP Platform Component and Connector View*

*SDP Operational System Component and Connector View*

# 7. References

## 7.1. Applicable Documents

The following documents are applicable to the extent stated herein. In the event of conflict between the contents of the applicable documents and this document, **the applicable documents** shall take precedence.
The list of applicable documents applies to the whole of the SDP Architecture.

| | |
|---|---|
| [AD01] | SKA-TEL-SKO-0000002 SKA1 System Baseline Design V2, Rev 03 |
| [AD02] | SKA-TEL-SKO-0000008 SKA1 Phase 1 System Requirement Specification, Rev 11 |
| [AD03] | SKA-TEL-SDP-0000033 SDP Requirements Specification and Compliance Matrix, Rev 02C |
| [AD04] | SKA-TEL-SKO-0000307 SKA1 Operational Concept Documents, Rev 02 |
| [AD05] | 000-000000-010 SKA1 Control System Guidelines, Rev 01 |
| [AD06] | 100-000000-002 SKA1 LOW SDP to CSP ICD, Rev 04A |
| [AD07] | 100-000000-025 SKA1 LOW SDP to SaDT ICD, Rev 04 |
| [AD08] | 100-000000-029 SKA1 LOW SDP to TM ICD, Rev 03B |
| [AD09] | 100-000000-033 SKA1 LOW SDP to LFAA Interface Control Document (ICD), Rev 01 |
| [AD10] | 300-000000-002 SKA1 MID SDP to CSP ICD, Rev 04A |
| [AD11] | 300-000000-025 SKA1 MID SDP to SaDT ICD, Rev 04 |
| [AD12] | 300-000000-029 SKA1 MID SDP to TM ICD, Rev 03B |
| [AD13] | SKA-TEL-SKO-0000484 SKA1 SDP to INFRA-AUS and SKA SA Interface Control Document, Rev 02 |
| [AD14] | SKA-TEL-SKO-0000661 Fundamental SKA Software and Hardware Description Language Standards |
| [AD15] | http://www.ivoa.net/documents/TAP/ |
| [AD16] | http://www.ivoa.net/documents/latest/SIA.html |
| [AD17] | http://www.ivoa.net/documents/DataLink/ |
| [AD18] | http://www.ivoa.net/documents/SSA/ |
| [AD19] | Memorandum of Understanding between the SKA organisation and National Radio Astronomy Observatory relating to a work package for the study and design of a new data model for the CASA software package |

[AD20] MeasurementSet definition version 3.0. MSv3 team, eds. 2018. http://casacore.github.io/casacore-notes/264

[AD22] Shibboleth Authentication Service from Interenet2 https://www.internet2.edu/products-services/trust-identity/shibboleth/

[AD23] COmanage Authorization Service from Interenet2 https://www.internet2.edu/products-services/trust-identity/comanage/

[AD24] SKA-TEL-SKO-0000990 SKA Software Verification and Testing Plan

## 7.2. Reference Documents

The following documents are referenced in this document. In the event of conflict between the contents of the referenced documents and this document, **this document** shall take precedence.

[RD01] SKA-TEL-SDP-0000046, SDP Costing Basis of Estimate

[RD02] SKA-TEL-SDP-0000115, SKA1 SDP RAM Analysis