Memo

To:	MeerKAT engineering team
From:	Jason Manley
Date:	15 January 2013
Memo no.:	!!DRAFT!!
Re:	MeerKAT data network

In this document, we outline the primary data network architecture options for MeerKAT. We aim to provide input on the type of cable to deploy for the AFN, RFN and KAPB internal data networks. In addition, we aim to select a long-life, low-cost networking solution for use within the KAPB.

Relative costs are considered for various network fabrics and transceivers suitable for achieving MeerKAT's core Ethernet switch requirements. Later sections attempt to estimate the number of ports required for each MeerKAT array release, and then make recommendations for cost-effective network scaling. Please note that these estimates are based on anticipated ROACH3 and GPU performance and actual processing performance may differ, resulting in a proportional increase or decrease in the number of processing nodes and switch ports.

We begin by introducing the current popular 1Gbps and 10Gbps Ethernet solutions and try to anticipate future industry direction for 40GbE and 100GbE.

Historical Ethernet

As of 2012, the commodity Ethernet standard is 1000BASE-T and is the current building wiring standard for small offices and home use, using Category 5e, 6 or 7 shielded or unshielded twisted pairs (four per link) and 8P8C ("RJ45") connectors. This copper cabling is ubiquitous, very cheap and can be terminated on site with low-cost crimping equipment. 1000BASE-T can drive network segment links of up to 100m. Most terminal equipment (computers) have support for 1000BASE-T using onboard Network Interface Cards (NICs). This technology has a very low total solution cost.

Because of this low cost, this solution is popular in smaller datacenters that need runs of less than 100m and low bandwidths. A popular arrangement is to place a large network switch at the end of a row of racks and directly run cables to each server in the row. End-of-row switches are then interconnected to a central switch. An alternative is to use small top-of-rack switches, with backhaul links to a large, centrally-located switch. This latter tree'd model works well when full-crossbar switching is not needed; 4:1 over-subscriptions of these backhauls are common (40x 1GbE ports to one 10GbE uplink, though some top-of-rack 1GbE switches support four 10GbE "uplink" ports).

To drive longer ranges, an aftermarket NIC or a switch that accepts SFP or GBIC modules is required. This provides the communication node with a port that can be populated with various optical transceiver modules that can drive up to 300m on OM3 Multi-Mode Fibre (MMF) or up to 80km on Single Mode Fibre (SMF) without any repeaters. The cost of these transceivers is significantly higher than the largely free 1000BASE-T ports.

Table 1: Differences between various common Ethernet links, ordered by link speed and range. Types available in 2012 are in black, with red denoting anticipated standards. Costs are indicative for a typical end-to-end node connection to a datacentre-grade switch port and includes a NIC, switch port and, where applicable, associated transceivers. Cabling itself is excluded and listed in the following column as a per-meter cost. In the case of fibres, this is for a single duplex link (fibre is cheaper in bundles, so this represents a worst-case cost).

Link	Medium	TNC Connector	Max Range	Fixed Cost (USD)	Cable cost (USD per m)
1000BASE-T	Cat5e 4pair Copper	8P8C	100m	165†	0.32
1000BASE-SX	50μ MMF pair	SFP	550m	388	1.62
1000BASE-LX	9μ SMF pair	SFP	10km	644	1.20
10GBASE-T	Copper Cat5e/6 U/UTP	8P8C	55m	742‡	0.32
10GBASE-T	Copper Cat6/6A/7 U/UTP or F/UTP or S/FTP	8P8C	100m	742‡	1.20
10GBASE-CR	Twinax copper	SFP+	7m	796	17.94
10GBASE-SRL	50μ MMF pair	SFP+	100m	1786	1.62
10GBASE-LRM	62.5μ MMF pair	SFP+	220m	3106	1.15
10GBASE-SR	50μ MMF pair	SFP+	300m	2050	1.62
10GBASE-LRL	9μ SMF 1310nm pair	SFP+	1km	2178	1.19
10GBASE-LR	9μ SMF 1310nm pair	SFP+	10km	3370	1.19
10GBASE-ER	9μ SMF 1310nm pair	SFP+	40km	17530	1.19
10GBASE-DWDM	9μ SMF 1540nm pair	SFP+	40km	15732	1.19
10GBASE-ER	9μ SMF 1310nm pair	SFP+	80km	20732	1.19
10GBASE-DWDM	9μ SMF 1540nm pair	SFP+	80km	20732	1.19
40GBASE-CR4	Twinax copper	QSFP+	7m	1416	55.38
40GBASE-AR4	Active optic cable	QSFP+	50m	1390	3.19
40GBASE-SR4	Parallel (8x) 50μ MMF	QSFP+	100m	5948	6.10
40GBASE-LR4	9μ SMF 1310nm pair	QSFP+	10km	21150	1.19
100GBASE-CR4	Twinax copper		3m		
100GBASE-AR4	Active optical cable		30m		
100GBASE-SR4	Parallel (8x) 50μ MMF		100m		6.10
100GBASE-SR10	Parallel (20x) 50μ MMF		100m		30.60
100GBASE-MR4	Parallel (8x) 9μ SMF 1310nm		1km		2.20
100GBASE-NR4	9μ SMF 1310nm pair		1km		1.19
100GBASE-LR1	9μ SMF 1310nm pair		1km		1.19
100GBASE-LR4	9μ SMF 1310nm pair		10km		1.19
100GBASE-ZR1	9μ SMF 1310nm pair		100km		1.19

†Assuming onboard NICs and native BASE-T switches. SFP 1000GBASE-T transceivers cost USD160ea.
‡Assuming native BASE-T switches and NICs. 10GBASE-T SFP+ transceivers do not exist due to power constraints of the SFP+ standard.

Faster links: 10GbE vs 40GbE vs 100GbE

Modern data centres are now deploying 10GbE en-masse, with 40GbE essentially reserved for backhauls. MeerKAT at L-band requires approximately 40Gbps from each antenna and so the 40Gbps Ethernet standards are well matched.

Link technologies

10GBASE-T is soon to be standard on server-class computers and is already offered as an option on some Dell and HP servers, for example. While already offering the lowest overall cost of all the 10GbE links, the commodity adoption of this standard will significantly further reduce the cost of this equipment and make 10GBASE-T links more attractive than 10GBASE-CR SFP+ direct-attach copper 'twinax' links. This is because 10GBASE-T cabling is backwards compatible with 10BASE-T, 100BASE-T and 1000BASE-T and also offers increased drive distances over a -CR solution. However, SFP+ connectors offer the most flexibility as they allow for copper or fibre links (albeit at increased cost), whereas BASE-T will always be limited to 100m.

While 1000BASE-T SFP modules are available, 10GBASE-T modules are not technically feasible due to increased 10GBASE-T PHY power requirements, which exceed the SFP+ specification. A mixed-deployment of SFP+ and BASE-T is thus unlikely as nodes would not be able to communicate without adaptors. 10GbE SFP+ copper (-CR or twinax) interfaces consume less power than the 10GBASE-T ports, as they do not contain the power-hungry BASE-T PHYs.

It is expected that add-in cards will be required for end-nodes if an SFP+ solution is adopted as we anticipate most commercial servers will ship with 100Mbps/1000Mbps/10Gbps BASE-T ports.

Scalability of the twisted-pair BASE-T standard beyond 10Gbps speeds is also in question. While 40Gbps speeds have been demonstrated, it remains to be seen if this solution can be produced cost-effectively for consumer deployment and no standard has been ratified or even proposed as of 2011.

For 40GbE deployments, QSFP+ is thus required. QSFP+ allows for backwards-compatibility with 10GbE using "spider" or "octopus" cables, which break-out a single 40Gbps QSFP+ port into four 10GbE SFP+ ports.

Datacenters

Topologies vary, but it is currently common practice in datacentres to deploy 10Gbps SFP+ direct-attached copper twinax to top-of-rack switches with 40bE fibre uplinks to an end-of-row (AKA aggregation) switch or else directly to the core switch in order to interconnect each rack. These top-of-rack switches typically have 40 10GbE ports and four 40GbE ports, realising a 2.5:1 over-subscription ratio. While this is sufficient for most compute/webserver type tasks, and for MeerKAT's CAM requirements, it is insufficient for MeerKAT's data transport purposes, where ratios approaching 1:1 are required.

Further, while small datacentres can manage with BASE-T connections and the associated 100m cable run limitation, larger datacentres can require longer cable runs, especially for links bridging rack rows to a distant core network switch. This, together with rising copper costs and routing issues of many copper cables (which have much larger outer diameters), is driving the increased use of fibre in larger datacentres.

Multimode fibre is currently the standard for these inter-rack or inter-row links within the datacentre. However, in light of industry trends towards SMF and upcoming silicon photonics technology (which favours 1510nm SMF), it can be argued that SMF deployment within the KAPB would be preferred in the interests of forwards compatibility.

Four fibre pairs are needed for 40Gbps MMF (40GBASE-SR4) solutions. For this reason, 40GBASE-SR4 eschews the normal LC connectors in favour of 12-fibre MTP/MPO (Multifibre-push-on), which are available

(a) MPO connector

(b) An LC fibre connector

Figure 1: A multifibre push-on (MPO) connector, as used for 40GbE MMF connections, contains a single row of 12 fibres, 8 of which are used per 40Gbps link (four in each direction). Looking forward, this connector will not provide a sufficient number of fibres to establish a 100GBASE-SR10 link, and two rows of fibres are needed. An LC fibre connector is already common in SMF links, which uses two uni-directional fibres to establish full duplex communication. The same LC connector is used for 100/1000/10G/40GBASE-LR standards, and will likely be used for the future 100GBASE-LR4 standard, providing simple forwards compatibility.

in pre-terminated patch lengths, with either only the needed eight fibres populated, or all 12 (with four unused) in loose-tube or ribbon cables. This increased fibre-count requirement raises costs for an MMF solution. Figure 1 shows such a connector.

In contrast, only a single fibre pair is needed per 40GbE SMF link (40GBASE-LR4). These typically employ the usual LC connectors, ensuring backwards compatibility with older standards. Coarse Wavelength Division Multiplexing (CWDM) is used on 40GBASE-LR4 and 100GBASE-LR4 standards to place four wavelengths on a single fibre. This makes them incompatible with many optics vendors' longhaul multiplexing systems which typically employ Dense Wavelength Division Multiplexing (DWDM). 40GbE uses four 10Gbps channels, which can be electrically (passively) demultiplexed into four discreet 10GbE channels and then DWDM multiplexed using existing longhaul equipment. 100GbE uses four 25Gbps channels.

Looking beyond 40GbE, 100GbE again requires only two pairs of SMF (100GBASE-LR4) but will need 10 pairs of MMF (100GBASE-SR10). 40GbE SMF deployments are then forwards compatible, whereas, were MMF to be deployed, datacentre upgrades would then either require modifications to the patch panels to support the larger 24 fibre MPO/MTP connectors, or else two existing 12 fibre connectors would need to be used.

Cisco has indicated that they consider 100GbE to be the next big step after 10GbE that will see large market penetration and that 40GbE is an interim transition. However, 40GbE has already seen adoption in larger datacentres, so it's not unreasonable to expect that the standard will be long-lived.

MeerKAT's L-band requirements are well matched to 40GbE speeds.

Scalable switches

A key requirement of the scalable architecture proposed for MeerKAT is a central full-crossbar non-blocking Ethernet switch, which needs multicast support. The largest commercial chassis units available off-the-shelf as at 2012 (such as the Arista 7508E at 288 40GbE ports) host less than the estimated 500 ports required for MeerKAT. For this reason, it is important to consider mechanisms for constructing larger switches.

It is possible to build larger switches using smaller units as building blocks. Charles Clos formalised this structure in the 1950s for the purposes of constructing circuit-switched telephone networks. Figure 2 demonstrates this basic architecture. For circuit-switched telephone networks, it was important that existing circuits not be interrupted when new calls are made. To ensure full-crossbar, non-blocking switching under these conditions, $m \geq 2n-1$ for a spine constructed from m switches of n ports.

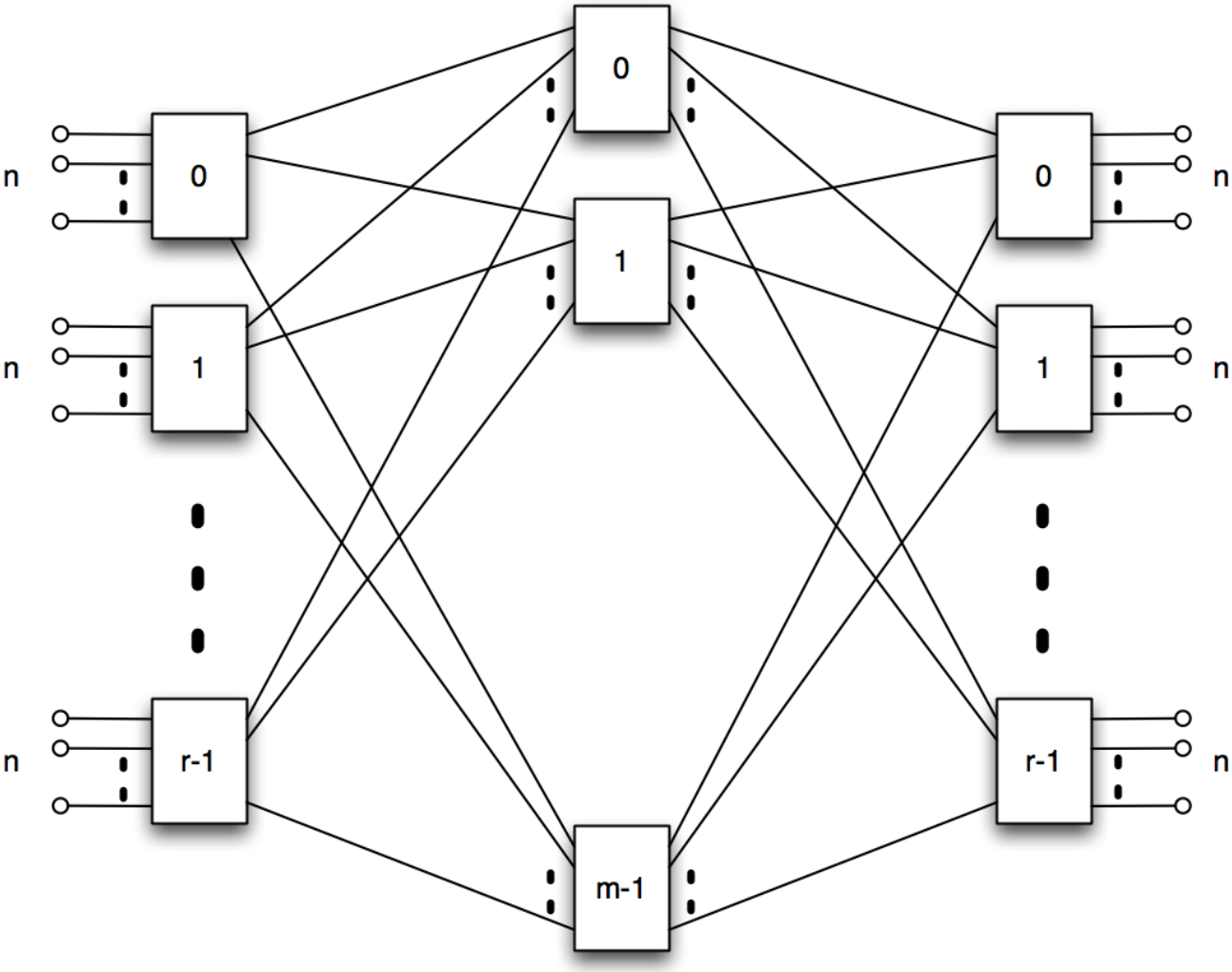

Figure 2: A Clos crossbar network switch, constructed from smaller crossbar switches. In a circuit-switched network, in order to ensure full crossbar, non-blocking capability, with the added proviso that existing connections no be rerouted, $m \geq 2n-1$. This figure shows a 3-stage Clos network, consisting of n ingress ports on each of the r ingress switches, m intermediate switches of $2r$ ports each, and n egress ports on each of the r switches in the egress stage.

For the purposes of bi-directional packetised computer networks, the Clos network can be folded to produce a *Fat Tree* architecture, as shown in Figure 3. In packetised networks, which can accommodate packets being

rerouted along a different route, the non-blocking requirement relaxes to $m \geq n$.

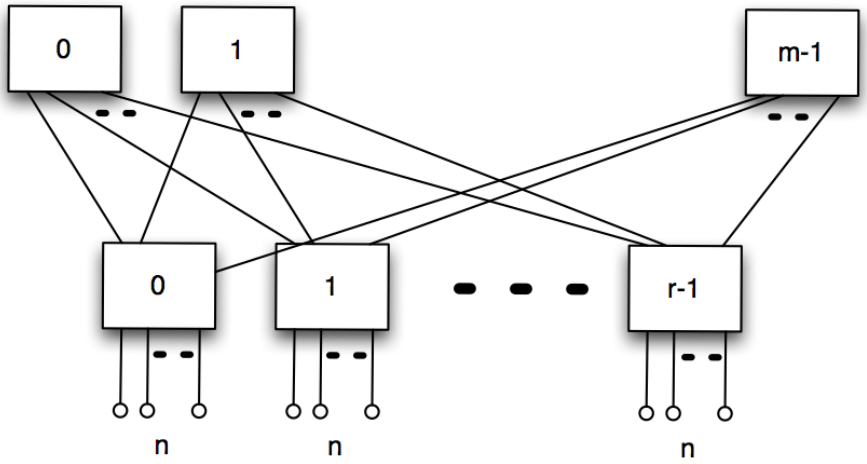

Figure 3: A folded Clos network, also known as a fat tree architecture.

The total number of switches in a folded 3-stage Clos network is then given by $\frac{2p}{s} + \frac{p}{s} = \frac{3p}{s}$, where $p \leq \frac{s^2}{2}$ for a total of p ports constructed from s port units.

When using practical 64-port switches as building blocks, a 3-stage Clos network will thus scale to 2048 ports, which is more than sufficient for MeerKAT's needs. Thereafter, 5, 7, 9 or any other odd number of stages (not detailed here) will allow further scaling to SKA sizes and beyond.

There are complications with the aforementioned scaling equation in keeping links to integer numbers. Consider trying to construct a 300 port switch from 32-port switches. The formulas above suggest that $\frac{3p}{s} = \frac{3\times300}{32} = 28$ switches are required. The ingress/egress layer would consist of $\frac{600}{32} = 18.75$ switches, so 19 are required. To interconnect these, $\frac{300}{32} = 9.375$ so 10 should be required. However, 10 doesn't divide into 19 cleanly. Whereas each ingress/egress switch should have 2 (1.9) links into the middle layer, only a single link to each of the 19 ingress/egress switches can be accommodated by each of the 10 middle-stage switches due to port count limitations (32, whereas $2 \times 19 = 38$ is required for two links). This means that 19 middle-layer switches are now also required, giving a total switch count of 38; a significant increase over the expected 28.

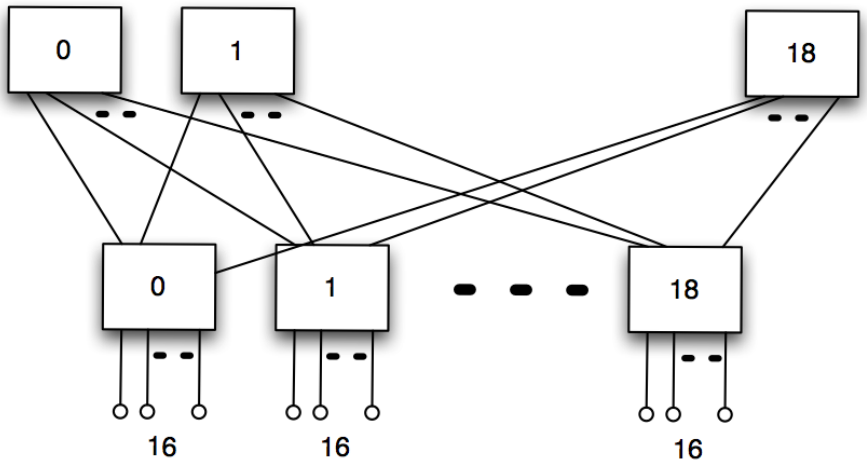

Figure 4: A 300 port Clos network constructed from 32 port units. The design does not achieve perfect utilisation: of the 32 ports on each middle layer switche, only 19 are used.

Routing traffic

In the Clos network, the ingress switches have multiple output ports through which packets destined for other egress switches can be routed. The non-blocking prediction whenever $m \geq n$ only ensures that there is a possible non-blocking route for packets from any ingress port to any egress port, but does not attend to the issue of how this route is determined.

How do ingress switches decide on which port to forward a given packet?

In the case of simple layer-2 Ethernet switches, each switch only has localised knowledge of port buffer statuses and capacities. It is thus impossible, without additional information, to intelligently select the end-to-end route with the lowest congestion.

Broadly, there are two options: randomly choose one, or deterministically decide which to use. If the traffic patterns are known, or can be predicted, then it is possible to statically configure specific routes based on source and destination without needing to know dynamic information about subsequent switches. Generally, however, this is not possible as computer network traffic patterns are not always predictable.

Common implementations are thus based on randomisation, such as the popular equal cost multi-path (ECMP) or the recently ratified Shortest Path Bridging (SPB). Over the Internet, different routes can exhibit different latencies and maximum transmission unit (MTU) differences. This can cause rapidly changing latencies and packet reordering if packets belonging to a single stream are distributed across different route paths. These effects can disrupt the operation of many Internet protocols, most notably TCP.

It is unlikely that MeerKAT will suffer from these particular problems for two reasons:

1. The Clos network will use the same or similar makes and models of switches which means that the fabric is uniform. MTUs will be the same and latencies through any intermediate switch (ie ports on different ingress/egress switches) should be similar.
2. MeerKAT will use stateless UDP, not TCP, which has no flow control mechanism like TCP whose back-off engine can be upset by changing latencies. The end-node receiver's buffer simply needs to be able to accommodate the difference between the maximum and minimum latencies through the network.

In an attempt to solve this problem in the general case, RFC 2992 analyses one particular multipath routing strategy involving the assignment of flows of data, rather than individual packets. The flows are identified by hashing flow-related data in the packet header (such as source and destination MAC addresses and IP addresses and ports), and this hash used to select an egress port. This allows for all packets from any particular network flow to traverse the network on a single path, while balancing multiple flows over multiple paths in general. This solution is commonly implemented on commercial switches.

In MeerKAT's case, then, the only significant consideration is that a routing path be chosen that will not result in the packet being dropped or significantly delayed (queued in an output buffer), which would adversely affect jitter and increase the buffer requirements in the processing nodes. Using a randomised approach, it is possible that suboptimal routes can be chosen, resulting in buffers filling. This is significant for MeerKAT, since each link is expected to be used at above 90% capacity, so individual links and buffers can easily become swamped. Since MeerKAT's traffic patterns are mostly deterministic, it should be possible to calculate what the minimum required port buffer size, given a certain routing algorithm. It is recommended that this be discussed with the selected switch vendor and a mechanism employed to allow even distribution of traffic over available fabric links.

Routing layer and practical limitations

In modern "layer2+" switches, packets can be routed using OSI layer 2 (MAC address) or layer 3 (IP address).

In general, a layer2-based system is not as scalable as layer3. Link Aggregation Group (LAG, IEEE 802.1AX-2008) allows for multiple Ethernet links to be grouped together into a single logical link, with all members sharing a single MAC address. While this allows a high-bandwidth point-to-point link, it does not address the distributed link case as required in a Clos network. Multichassis link aggregation group (MC-LAG) is often implemented by vendors, but this is not covered under IEEE 802.1AX-2008 LAG standard. MC-LAG allows so-called inverse multiplexing of multiple ethernet links.

The IS-IS link state protocol (which runs on top of layer-2/Data link layer) is often used to compute the available route costs using Dijkstra's algorithm. This is used in layer-3 routing schemes, where multiple protocols are also available, the two most popular being TRILL (IETF standard) and SPB (IEEE 802.1aq standard). Both fundamentally are able to perform the function of Spanning Tree over multiple links, allowing effective use of available bandwidth.

However, waters become murky with few options when multicasting is required over these links.

WIP – explain differences of packet walks!!!!

(S,G) vs (*,G) entries. PIM-Sparse.

KAT-7 uses IP addresses for application-layer routing, and so places restrictions on the IP addresses of each node. It is possible to select addresses that meet both the requirements of layer 3 routing and application layer routing.

This problem is also present in multicast networks, where nodes across multiple layer-3 network segments subscribe to the same multicast group. In the multicast environment, the switch will need to maintain a routing table and layer-3 routing rules are generally not possible. Current multicasting support, routing strategies and vendor implementations on commercial layer-2+ switches in a Clos network environment is unclear.

Receptor fibre network (RFN) fibre selection

Single Mode Fibre (SMF) will be required for the Timing and Frequency Reference (TFR) and so there will already be SMF in the Receptor Fibre Network (RFN) bundle between the feed assembly and the pedestal. As of 2012, it will be more cost effective to deploy 10GbE on the pedestals, with parallel MMF employed to achieve higher datarates. The pedestal-based switch can then be of the top-of-rack 10GbE/40GbE variety offering many 10GbE ports and some 40GbE ports for uplink (such as the Arista 7150S-64 or the Cisco 3064X, each with 48 10/1GbE ports and 4 QSFP+ ports).

From a maintenance perspective, an optimal solution would be to deploy G652D SMF throughout the RFN and use four 10GBASE-LRL links to the L-, UHF- and S-band digitisers for data transport, with 16 for the X-band. Using SMF everywhere would allow for simple fibre exchange and transceiver sharing in the event of a failure.

Further, it is recommended that LC connectors be deployed in the pedestal patch panel for interfacing with industry-standard networking equipment and to allow the use of popular, pre-made, commercial LC patch cables.

Long range fibre choice (AFN) fibre selection

Between antennas, the installation of SMF is the only sensible choice, due to required transmission distances.

MeerKAT's antennas need to transport a bandwidth of approximately 40Gbps over a distance 12km for L-band, and closer to 160Gbps for X-band. Considering bandwidths, 40GBASE-LR4 seems appropriate. Considering the link budgets, 12km transmission should be possible with the 10km -LR4 standard and there are reports of

successfully establishing reliable links at over 20km using 10GBASE-LR transceivers. IEEE defines such links as “engineered links”. Should these prove inappropriate, longer distances can be driven using more expensive -ER transceivers.

From Table 1, it is evident that in 2012 it would be cheaper to deploy multiple 10GbE links rather than 40GbE. This is because 10GbE is already a mature product with greater market penetration than the newer 40GbE standards. 40GbE might become financially viable within MeerKAT’s timeframe, but it is difficult to predict what market choices will be.

Fortunately, the choice of fibre can be made independently of the Ethernet standard. 10GBASE-LR, 40GBASE-LR4 and 100GBASE-LR4 will all use the same SMF at wavelengths of 1310nm. For this reason, we recommend the installation of G652D SMF in the AFN.

Further, it is likely that 1550nm will see greater uptake in future, especially in light of silicon photonics, which favours longer wavelengths for lower losses. Silicon photonics transceivers are expected to be commercially available in 2014. For this reason, it is recommended that, if possible, installed fibres attempt to maintain compatibility with 1550nm windows.

KAPB CAM network cabling selection and switch topology

The standard 1Gbps copper cable will be sufficient for control and monitoring (CAM) purposes. Due to the low cost and the fact that these ports are standard on all compute servers, we recommend 1000BASE-T interfaces for all CAM nodes.

To accommodate interconnection on the core 10Gbps KAPB switch, we recommend a tree’d architecture and the use of Top-of-rack (TOR) 10GbE to 1GbE breakout switches. Oversubscription is possible with CAM links, since the datarates are anticipated to be very low. Since the KAPB switch is using SFP+ ports and the distances to each rack will exceed the 7m maximum SFP+ direct-attach range, a fibre solution will be required. In light of the similar pricing of -SR MMF and -LRL SMF solutions, it is recommended that these links should be SMF, which will also provide forwards compatibility with emerging technologies.

KAPB data network cabling selection and switch topology

Cabling

Copper interconnects in the form of 1000BASE-T on category 5 or category 6 cabling will be insufficient for MeerKAT’s data network. MeerKAT will need interconnect speeds well in excess of 10Gbps per processing node for data transmission. BASE-T standards are only defined up to 10Gbps speeds, with the future of faster copper standards uncertain. For this reason, we cannot recommend the deployment of category 5, 6 or 7 twisted pair cabling in the KAPB for data transmission purposes.

Shorter range “direct attach” standards, are available at 40Gbps speeds (QSFP+: 40GBASE-CR4), but only for runs up to 7m. This is an option for cabling within a rack only. It is not clear if the QSFP+ connectors will be compatible with 100GbE and future Ethernet standards, but it is highly unlikely that 40GbE direct attach copper cables will be forwards compatible. For this reason, we recommend limiting the use of direct-attach copper to short-lived applications, such as within a rack to processing nodes, which are expected to age quickly.

For KAPB inter-rack links, it seems prudent to use the fastest speeds possible as this will reduce cabling requirements and hence increase reliability. Assuming 40GbE speeds, these links will necessarily have to be fibre-based, due to link distances. They could take the form of SMF or MMF. It is possible that all networking

equipment will see limited reuse and that any future upgrades to MeerKAT's core switching infrastructure in future will require replacement of switches and rack cabling.

It would be ideal to try to limit any future recabling. We should try to limit the recabling requirements to rack-level only, and attempt to maintain forwards compatibility for inter-rack links, to avoid disturbances to existing infrastructure during future upgrades. SMF would provide such forwards compatibility, whereas MMF would not (additional MMF fibre would need to be installed to support faster Ethernet links). Since there is no cost advantage for MMF over short-range SMF, we would recommend using SMF for inter-rack links within the KAPB.

However, Active Optical Cables in lengths up to 50m are also available for 10GbE and 40GbE at significantly lower costs than discreet transceivers and fibre. If cost becomes a significant factor, these cables would allow for cost-effective inter-switch links across racks but would offer no forwards compatibility with future Ethernet standards.

Switch topology

We will now outline three possible options for the KAPB data network topology:

- banks of 1- or 2-RU top-of-rack (TOR) switches distributed throughout the compute racks, or,
- a combination of TOR and chassis switches, or,
- regularly-spaced unified, large chassis switches.

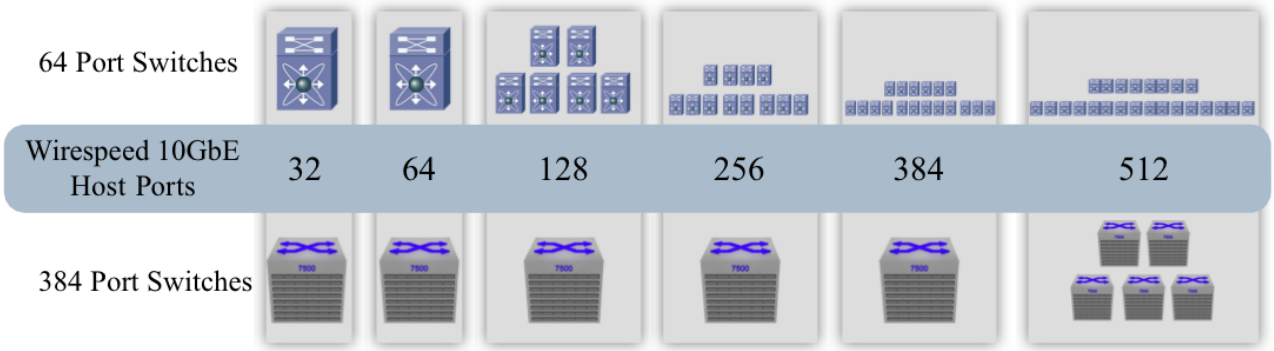

Figure 5: Scaling switches using 1-RU vs chassis switches, courtesy of Arista, from `http://www.aristanetworks.com/media/system/pdf/Network_Scalability_AAG.pdf`.

The switch requirements for these two concepts is illustrated in Figure 5, showing the Clos scaling using chassis and 64-port 1-RU units for the 10GbE case using 2011-model 10GbE switches. Both techniques will now be discussed in the following two sections.

Indicative pricing from two vendors has revealed that chassis-based solutions are significantly more expensive than solutions based on TOR switches.

Clos network of top of rack switches

In this topology, each rack contains a small (1-RU or 2-RU) switch. This 'leaf-node' (aka 'access layer') switch connects all devices within the rack using some of the available ports. The other ports are used to connect to

adjacent racks via a second tier of 1- or 2-RU switches, forming a 'spine' (aka 'aggregation') layer to form a 3-stage Clos network[1] sufficient for MeerKAT.

This solution allows for 10GbE, 40GbE or a mixture of both or any other link speed, with per rack selection of ports, depending on the selection of top-of-rack (TOR) leaf node switches. However, 10GBASE-T switches with QSFP+ upload ports typically oversubscribe uplinks too much for MeerKAT's purposes (ratios of 48x 10GbE to 4x40GbE are common). This limitation can be overcome by supplementing the uplink ports using the ordinary 10GbE ports and 40GbE octopus/spider/breakout cables.

Most TOR switches have 24, 32, 48 or 64 ports. This solution would thus allow for up to 32 nodes per rack while still providing a 1:1 contention ratio on all ports. MeerKAT can accommodate some limited oversubscription on certain components and so it is possible that this number can be increased to at least 36 without compromising performance for all engines except the D- and F-engines, which will need to be hosted on true 1:1 contention 40GbE ports.

Alternatively, a pure 40GbE solution can host 'spider' breakout cables for 10GbE connections, which converts each 40GbE port into four 10GbE ports. Depending on the manufacturer and implementation, hardware limitations can restrict the number of simultaneous 10GbE ports. So care must be taken if this approach is adopted. This is also only possible using 40GBASE-SR4 (expensive) or 40GBASE-CR (short range) cables.

A TOR solution would necessarily use SFP+ or QSFP+ switches, which are capable of hosting optical modules or active cables, since each rack must interconnect to racks up to 15m away. Direct-attach links are capable of linking all intra-rack compute nodes to the TOR switch.

Item	Est. per-unit cost	Qty	Total
64 port 40GbE leaf nodes	USD70000	16	USD 1.1M
64 port 40GbE spine nodes	USD70000	8	USD 560k
40GBASE-CR4 client links	USD270	512	USD 138k
40GBASE-AR4 leaf-spine links	USD390	512	USD 200k
Total			USD 2M

Table 2: The estimated cost for constructing a 512 port Clos network based on 64 port switches, with 1:1 contention ratios using low-cost AOC cables for inter-switch links. Leaf nodes are able to use low-cost direct-attach copper links as clients are located in the same rack as the switch.

TOR Clos Advantages:

- Lowest cost.
- Modular design allows for incremental buildout and mixture of switch technologies.
- Tight control of oversubscription.
- N+1 redundancy possible.
- Contained copper cabling (within a rack) with fibre across racks reduces overhead tray loading.

TOR Clos Disadvantages:

- Unable to mix 10GBASE-T, necessitating the use of SFP+ cards in servers.
- Smaller switching buffers of 1RU switches not tolerant of all traffic patterns.

[1] `http://en.wikipedia.org/wiki/Clos_network`

Combined top of rack and chassis switches

This is a variant of the afore-mentioned pure TOR solution that replaces the spine layer switches with fewer, chassis-based units. These chassis can be located centrally, and can be interconnected with low-cost copper cabling.

Item	Est. per-unit cost	Qty	Total
64 port 40GbE leaf nodes	USD70k	16	USD 1.1M
256 port 40GbE chassis	USD2160k	2	USD 4.3M
40GBASE-CR4 client links	USD270	512	USD 138k
40GBASE-AR4 leaf-spine links	USD390	512	USD 200k
Total			USD 5.7M

Table 3: The estimated cost for constructing a 512 port switch from TOR leaves and chassis spines, with 1:1 contention ratios.

Mixed TOR — chassis Advantages:

- Large spine chassis switches offer larger buffer sizes, potentially helping to level bursty traffic patterns.

Mixed TOR — chassis Disadvantages:

- Unable to mix 10GBASE-T and (Q)SFP+ on leaf nodes, necessitating the use of more expensive add-on NIC cards in servers.
- Larger points-of-failure in chassis.
- Smaller switching buffers of 1RU switches not tolerant of all traffic patterns.
- Increased costs.

Chassis-based end-of-row switching

This topology places a large switch (modular, chassis-based) at the end of each row of racks (or serves a group of racks). Copper or fibre connections run directly from this switch to each processing node in that row (or group) of racks.

It would be cost effective to use 10GBASE-T links in this solution; but this presents scalability issues as the future of copper cabling within the datacentre is uncertain. Also, it hampers scaling by consuming additional addressing table space: four MAC and IP addresses are required for each MeerKAT processing node in order to support 40Gbps speeds.

Item	Estimated cost	Qty	Total
288-port 40GbE chassis	awaiting pricing	6	USD ???
XX-port 40GbE linecards	awaiting pricing	???	USD ???
10GBASE-T Cabling	BASE-T linecard available?		
40GBASE-SR4 transceivers	USD2400	1050	USD 2520000
Total			**USD**

Table 4: The estimated cost for constructing MeerKAT's DBE using chassis switches.

Advantages:

- Lower total cost with fewer cable links.
- Link speed selection and flexibility through linecard exchange.
- Larger switching buffers tolerant of arbitrary traffic patterns.
- Mixture of 10GBASE-T and 10GBASE-CR, 40GBASE-CR or optical modules possible.

Disadvantages:

- Will require large switch investment to switch to faster Ethernet speeds.
- Coarse or single point(s) of failure.
- Incremental upgradability hampered by chassis.

Multicasting requirements

Function	Multicast Groups
Raw digitised stream	64
Transient buffer	64
Antenna coherency products	64
Wideband, 64 F-engs + 64 X/B-engs	
Channelised baseband	64
Beamformer (4 full time resolution beams)	256
Beamformer (100 time averaged beams)^2	1280
Correlator	64
Spectral line, 64 F-engs + 5 X/B-engs	
Channelised baseband (5 subbands)	5
Correlator	5
Total	**1866**

Table 5: The estimated number of multicast groups required by the full DBE, assuming 64 frequency channel groups (giving 13.375MHz resolution for L-band) over the full digitised bandwidth.

Anticipated scaling

MeerKAT will be constructed in stages, and it is expected that the DBE and hence networking infrastructure must also scale with the array. Figure 6 shows the block diagram functions required of MeerKAT's DBE. Figure 7 illustrates the dataflow through the switch for the correlator and beamformer components of this machine. Please note that these estimates are based on anticipated ROACH3 and 2014 GPU performance and actual processing performance may differ, resulting in a proportional increase or decrease in the number of processing nodes and switch ports.

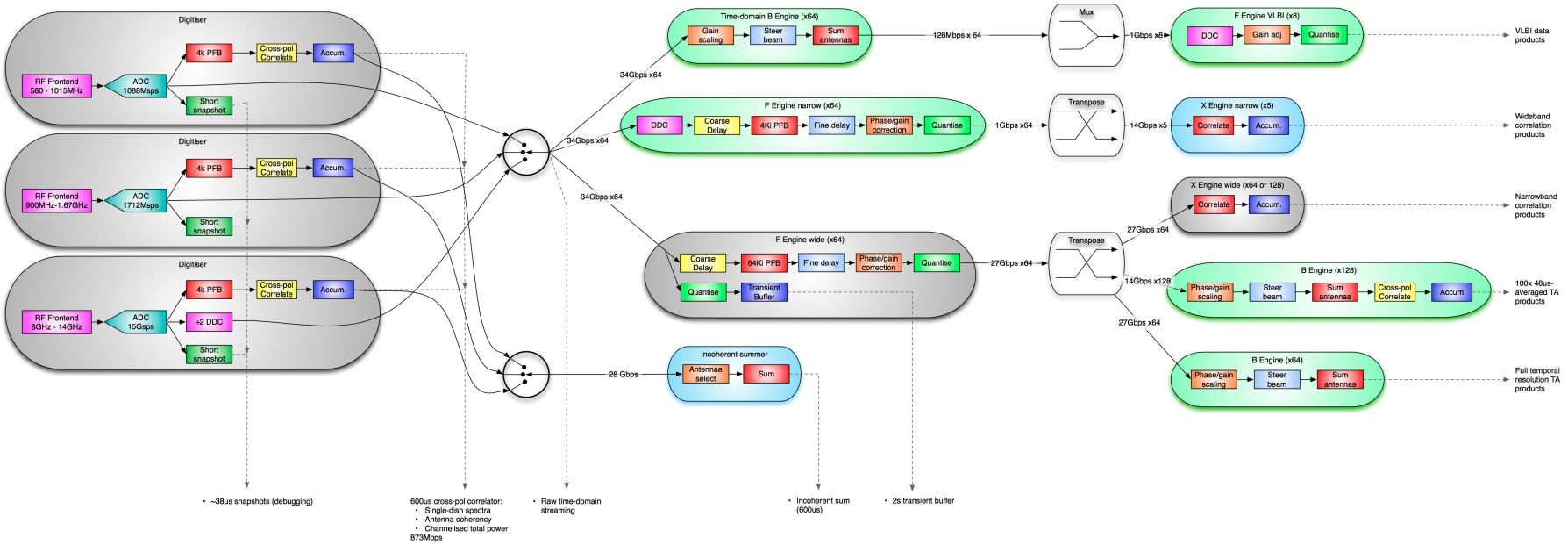

Figure 6: The MeerKAT DBE design has not yet been finalised, but this diagram shows a possible signal processing pipeline and associated datarates between engines.

Early prototype testing

Two dishes will be available for testing and verification during the first half of 2014. This system must be ready to accept 8 antennas by the end of 2014, with wideband correlation and a boresight beamformer functions. It is likely that a single 32-port switch will suffice for this purpose, with all equipment housed in a single rack. In all likelihood, this will be based on ROACH-2 hardware since ROACH-3 will not yet be available. The correlator and beamformer functions will be ported from KAT-7, with familiar user interfaces.

Table 6: **The number of switch ports needed for the initial 2-dish verification backend, assuming 40GbE interconnect.**

Description	40GbE ports
D-engines	2
F-engines	8
X-engines	8
CAM interfaces	8
40GbE output stream (SPT)	2
Total	**28**

Array Release 1

Array Release 1 (AR-1) will start with more than 4 antennas in early 2015, with scaling up to 16 antennas during the year. This system needs to support wideband correlation and beamforming for the 16 dual-polarisation inputs, which will require additional hardware and increase the system size beyond a single rack with a single switch.

Table 7: **Array release 1 is larger than the preceding testing correlator and adds beamforming for up to 16 inputs.**

Description	40GbE ports
40GbE links from each dish for L-band	16
Wideband F-engines	16
Wideband X-engines	16
Beamformers	16
CAM interfaces	2
40GbE output stream (SPT)	2
Total	**68**

Array Release 2

Wideband correlation, narrowband correlation, incoherent sum of antennas channelised power, Fly's eye, pulsar searching and subarraying modes for up to 32 dual-pol inputs.

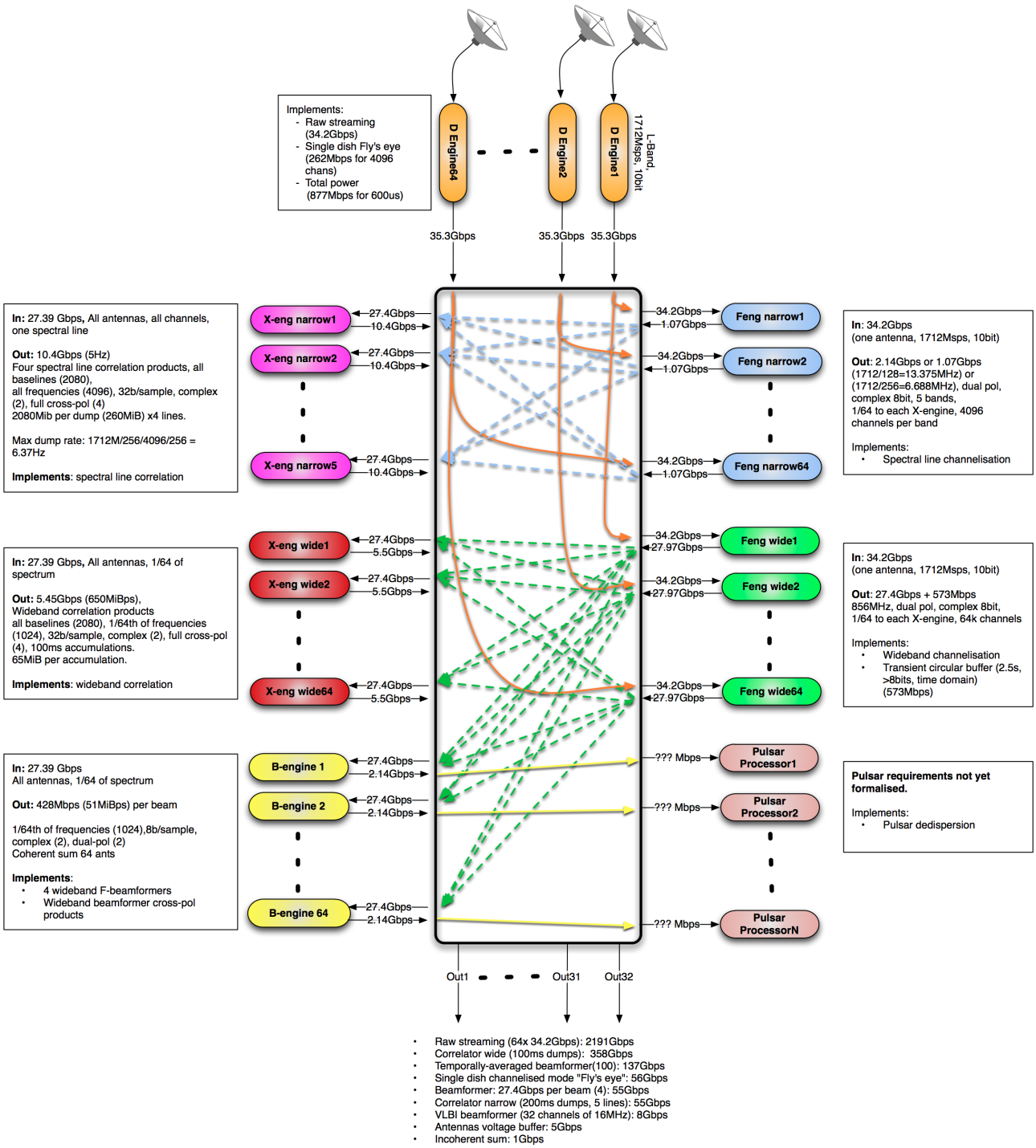

Figure 7: The anticipated L-band switch traffic for the MeerKAT instrument, should all instruments be used concurrently. However, the beamformer and narrowband (spectral line) correlators, for example, are not required simultaneously and are likely to share hardware.

Table 8: **Array release 2 requires additional functionality and processes more inputs than AR-1 and so also requires significantly more processing hardware and interconnect. This system copes with up to 32 inputs.**

Description	40GbE ports
40GbE links from each dish for L-band	32
Wideband F-engines	32
Narrowband F-engines	32
Wideband X-engines	32
Narrowband X-engines	5
Beamformers	32
Incoherent summer	1
CAM interfaces	2
40GbE output stream (SPT)	8
Total	**84**

Array Release 3

Wideband correlation, narrowband correlation, incoherent sum of antennas channelised power, Fly's eye, pulsar searching and subarraying modes for up to 64 dual-pol inputs.

Table 9: **Array release 3 is designed to process up to 64 inputs, with all features supported.**

Description	40GbE ports
40GbE links from each dish for L-band	64
Wideband F-engines	64
Narrowband F-engines	64
Wideband X-engines	64
Narrowband X-engines	5
Beamformers	256
Incoherent summer	1
CAM interfaces	2
40GbE output stream (SPT)	**16**
Total	536

Required output links

Table 10 shows the defined output and intermediate data products for MeerKAT AR-3 along with their datarates. From this table, the maximum total L-band DBE to SPT datarate for each mode can now be calculated for the various combinations of AR-3 instrumentation.

Imaging mode: 290 Gbps for the following output products...

- Interferometric visibilities WHR

Instrument	Maximum datarate
Raw digitised stream[a]	2191 Gbps
Channelised wideband[b]	1753 Gbps
Interferometric visibilities WHR [c]	175 Gbps
Temporally-averaged beamformer[d]	137 Gbps
Channelised narrowband[e]	68 Gbps
Single-dish antennas coherency products[f]	56 Gbps
Coherent tied-array beams[g]	55 Gbps
Interferometric visibilities NBR [h]	55 Gbps
Interferometric visibilities NCR [i]	55 Gbps
Interferometric visibilities WLR [j]	22 Gbps
VLBI beamformer[k]	8 Gbps
Antennas voltage buffer[l]	5 Gbps
Interferometric visibilities WTR [m]	3 Gbps
Channelised incoherent sum[n]	1 Gbps

Table 10: **MeerKAT L-band datarates, excluding packetisation and protocol overhead. This realtime machine will need to switch nearly 5Tbps during normal operations.**

[a] 128 streams, 1712Msps, 10-bit

[b] 8-bit complex per sample

[c] Wideband, high resolution (a.k.a. "Sa"), using a 100ms accumulation period, 32768 channels, 32-bit complex output, 4 cross-pol terms, 2080 baselines.

[d] 100 beams, 4096 channels, 47.85μs averaging, 8-bit complex

[e] 5 bands of 13.375MHz, 8-bit complex per sample

[f] Auto-correlations, incl. cross-pol terms, 598.13μs accumulations, 4096 channels, 32-bit

[g] 856MHz, 8b complex, 4 beams

[h] Narrowband B (aka "Sb"), using 200ms accumulation periods, 5 spectral lines, 4096 channels each, 32-bit complex, 4 cross-pol terms, 2080 baselines.

[i] Narrowband C (aka "Sc"), using 200ms accumulation periods, 5 spectral lines, 4096 channels each, 32-bit complex, 4 cross-pol terms, 2080 baselines.

[j] Wideband low resolution, using a 100ms accumulation period, 4096 channels, 32-bit complex output, 4 cross-pol terms, 2080 baselines.

[k] 32 channels of up to 16MHz each, 8-bit complex

[l] 2048MiB, 128 buffers, 1/60Hz readouts, 8-bit real

[m] Wideband, transient resolution: 100ms accumulation periods, 512 channels, 32-bit complex, 4 cross-pol terms, 2080 baselines.

[n] 598.13μs accumulations, 4096 channels, 32-bit complex, dual polarisation

- Interferometric visibilities N*R
- Incoherent sum of antennas' channelised power
- Antennas coherency products
- Antennas voltage buffer

The maximum output datarate is found in this mode. It is dominated by the 65k channel high resolution wideband (WHR) mode. Practically, it is likely that only 4096 channels (WLR) will be required for many observations, which will reduce the datarate to under **139Gbps**.

Pulsar timing: 120Gbps for the following output products...

- Interferometric visibilities WTR
- Coherent tied-array beams
- Incoherent sum of antennas channelised power
- Antennas coherency products
- Antennas voltage buffer

Transient search: 276 Gbps for the following output products...

- Interferometric visibilities WLR,
- Temporally-averaged channelised tied-array
- Incoherent sum of antennas' channelised power and
- Antennas voltage buffer
- Antennas coherency products
- four tied-array beams,

Fly's eye: 62 Gbps for the following output products...

- Incoherent sum of antennas channelised power
- Antennas voltage buffer
- Antennas coherency products

VLBI: 70 Gbps for the following output products...

- VLBI beamformers
- Incoherent sum of antennas channelised power
- Antennas voltage buffer
- Antennas coherency products

Conclusion

It is recommended that the RFN delay the choice of fibres until the last possible moment. If short-range SMF transceivers (10GBASE-LRL) become available within 2013, deployment of G652D is recommended for datalinks. Should these products not become available at a competitive pricepoint, 10GBASE-SR over OM4 is recommended for the data links. In either case, LC connectors will be used by the switching equipment.

The AFN should deploy G652D fibre for 40GBASE-LR4 links to each dish, to operate in the 1310nm window. Initially, only a single fibre pair will be used for UHF and L-bands, with three additional pairs reserved for increased X-band datarates.

AR-1 should deploy a single 1- or 2-RU switch for interconnect, and be housed in a single rack with QSFP+ direct-attach cabling to processing elements. Further, SMF should be deployed wherever possible, rather than MMF, for intra-system links.

AR-2 will require a multi-rack solution ($\sim$5 racks). If short-range SMF transceivers become available within the AR-2 design timeframe, it is recommended that SMF be deployed throughout the KAPB for inter-rack links, along with LC patch-panel connectors at each rack. If these devices do not materialise at MMF-competitive pricepoints, MMF is recommended with MPO connectors.

AR-2 can choose to extend the single switch from AR-1 by augmenting it with additional 1-RU units to form a 96 port switch, or, to discard this switch and to purchase a single chassis-based switch instead. Initially, this would only be partially populated with linecards. The decision for chassis-based or 1-RU-based switching topology should be made at this time to avoid costly replacements and enable equipment reuse for AR-3 deployment. It is recommended that a switch study be conducted to evaluate the various routing protocols for multicast performance. A Clos-based network of smaller switches is a more cost-effective solution than a chassis-based solution.

AR-3 will require 16 racks, with significant fibre interconnect between the racks. The AR-2 network should be extended, rather than a replacement network deployed. 16 40GbE connections are budgeted for SPT interconnect, which is sufficient to transport all required output dataproducts. However, should raw streaming be desired in future, 64 links will be needed. It would be useful to deploy spare ports to enable user-supplied-equipment (USE) to be connected for specialised science observations.