MULTI-TERABIT ROUTING IN THE LOFAR SIGNAL AND DATA TRANSPORT NETWORKS

Jaap D. Bregman , Gideon W. Kant, Haitao Ou

ASTRON, Oude Hoogeveensedijk 4, 7991 PD, Dwingeloo, The Netherlands

Email: {bregman, kant, haitao}@astron.nl

ABSTRACT

The LOFAR telescope is being developed as a giant data processing machine for astronomy. An assessment is presented of photonic technologies that will enable LOFAR to be materialized in the timeframe after 2004 and we discuss the basic cost trade-offs in photonic signal transport. A summary of the LOFAR configuration and architecture is given that provides the proper input for the cost equation and leads to the definition of a short and a long-range network. For interfacing the antenna clusters as well as the processing nodes to the network we propose ten Gigabit Ethernet routing technology being developed for the PC server market.

INTRODUCTION

The digitised signals from over ten thousand antennas are combined to produce data for astronomical images, which requires an aggregated data transport bandwidth of order twenty Terabits per second in the LOFAR system [1]. There are two kinds of signal transport networks each providing half of the capacity, one for the remote antenna stations and another for all the antennas within about ten kilometre from the central processing system. The long-range network will be build based on next generation wavelength division multiplex (WDM) or optical time division multiplex (OTDM) telecom technologies, which are investigated in the Retina project [2]. An assessment study shows that the short-range network could be realized in Ten Gigabit Ethernet transport (10GbE) technology being developed for the PC server market and becoming commercially available in the coming few years. In the paper we discuss potential implementation options and use numbers with the proper order of magnitude that adhere to the basic characteristics of LOFAR as presented in the architectural design document [3]. An important aspect to be discussed is the routing configuration that interconnects order two thousand receive cluster nodes and order five hundred processing nodes to the array signal network through 10GbE ports. At the central processing facility we have a data network that interconnects order two thousand processing nodes that perform the array cross-correlation between the station beams or performs a multi beam forming operation that images the full sky. The latter operation requires twelve Terabit per second routing capacity from receive clusters to processing nodes for which a butterfly configuration of multi port router devices is proposed.

LOFAR CONFIGURATION

LOFAR is an aperture synthesis array with over hundred antenna stations, each with order hundred dual polarization receptors. The stations have an exponentially increasing distance from the centre and follow curved arms that extend over 200 km. About a quarter of the stations have more or less random positions within a circle of 1 km radius leaving fifteen stations on each of the five arms. In Fig. 1 the station distribution along one such arm is indicated. So half of all stations are within 6 km of the centre and the last increment is about 60 km. We assume a central data processing facility within 3 km of the array centre.

| .4 | .6 | .8 | 1.2 | 2 | 3 | 4 | 5 | 8 | 10 | 15 | 20 | 30 | 40 | 60 |

Fig. 1. Increment in km between 15 stations (|) along an exponential spiral arm (not to scale).

PHOTONIC ASSESSMENT

State-of-the-art digital signal transport uses multiple 3 Gb/s electrical serial links over transmission lines and 10 Gb/s serial links over optical wave-guides. Conversion chips from low speed parallel data into high-speed serial format are priced in the tens of Euro range and allow cost effective replacement of heavy parallel connectors and cables by serial ones. Electrical serial cables, although more expensive than fibre optical ones, are cost effective at short distance (ten metre) when the price difference between electronic (tens of Euro) and optical (hundreds of Euro) transmission line drivers and receivers is taken into account. Once in the optical domain, cable attenuation is so low that distances op to tens of kilometres can be bridged without further equipment. A bare fibre price of 0.1 €/m then indicates that a few kilometres can be bridged before fibre cost starts dominating the total link cost.

Adding a photonic transceiver only doubles the price of a Gigabit Ethernet over twisted pair interface card used to interconnect PC type servers in a network. In 2002 all the key components to build 10GbE interface cards are commercially available, and we expect complete products by the time LOFAR equipment needs to be installed at prices that are then only a few times the current price of 1000BASE-SX Gigabit Ethernet over fibre equipment. Even router chips with over hundred I/O channels and a throughput of 0.5 Tb/s are available for a few thousand Euros, which indicates that the cost of router boxes will be dominated by the data interfaces. The PC server based network applications organized in Wide and Metropolitan Area Networks are expected to use 10GbE technology and constitutes a market big enough to warrant PC level prices for this high speed data communication.

NETWORK COST EQUATION

The total cost of a connection between a LOFAR station and the central processing facility includes the cost of trenching, the number of fibres in a trench and the photonic transceiver. We neglect the cost of pre-processing to reduce the data rate. In designing a system we combine technologies that have different marginal cost effects when the system is rescaled in a certain parameter. For a data transport network the performance issues are distance and bandwidth. An optimum choice is found when all three cost contributions are about equal. Fibre and transceiver cost are matched at an average fibre length of three kilometres. For a trunk arm with an exponentially increasing distance between the branching points we then find a maximum distance of about six kilometres from the centre to provide an average fibre length of three kilometres. Then half of all receptors could be connected to the central processing centre. Transporting the full potential bandwidth requires one fibre at 10 Gb/s for every five dual polarization receptors, which leads to a total of thousand fibres with an average length of 4.5 km to the processing centre. The five arms then need thirty kilometres of trenching, which means an average of hundred and fifty fibre pairs in a trench. Typical trenching cost is tens of Euros per metre and just equals the fibre cost. Having made these initial choices the cost of trenching is a third of the total cost. The average trench length for the fifty stations of the inner array is now 0.6 km per station and we can express the average station connection cost as just two kilometres trench equivalent. The marginal cost of adding one more station on an arm is dominated by the trenching cost, the so-called last mile problem. However connecting the next two stations on each arm requires just over three kilometres per station and does not yet increase the average cost.

We have shown that full 200 Gb/s data transport from the centre 60 % of the stations out till thirteen kilometres from the centre is a system optimum where the cost of trenching, fibres and transceiver is balanced.

Connecting the more remote stations by exponentially extending the trench along the arms drives up the average connection cost. This means that we have to look into alternative solutions that share the dominating trenching cost with other users. The most attractive route is leasing dark fibre from external parties and use own equipment. The fibre and shared trench cost for these remote stations dominates over the cost of the 10GbE transceiver boards, and it becomes cost effective to use WDM or OTDM technology to transmit 40, 80 or 160 Gb/s from each remote station instead of only 10 Gb/s on a single fibre.

This shows that the remote stations can operate at 10 Gb/s in the initial phase of the project requiring forms of data compression and reduction at the remote stations. When the advanced data transport technology becomes available in a later phase, it could be installed together with additional processing hardware to support the enhanced bandwidth of the system.

LOFAR ARCHITECTURE

LOFAR is a giant data processing machine where signal data from a few thousand receive clusters are pre-processed and transported to a central facility for further processing. A basic logical block is the receive cluster where the signals

of about five dual polarization receptors are digitised and buffered before asynchronous transport to the processing nodes where the beam forming operation is executed. Further architectural elements are routing devices, fibre optical cabling network, optical repeaters and fibre optical multiplex units. It is attractive to use a single type of data interface for all this equipment and 10GbE is proposed as being well matched to commercial off-the-shelf equipment in the 2005 time frame. The central facility is a cluster of a few thousand PC servers interconnected with a high bandwidth network that transports the signals between all the processing nodes.

Three processing operations are performed on the antenna signals. First the wide band signal of each antenna is divided in a number of sub bands. The second operation is beam forming, where the signals of a group of clusters that constitute a station are combined for each sub band into a beam that selects a certain part of the sky. Finally the beam of each station is cross-correlated against the beam of all other stations. The second and third class of operations require each about thousand processing nodes. Therefore, frequency or time slicing needs to be done on the receptor data streams such that the slices are routed to different processing nodes. Then the data slices from different receptors are merged for further combined processing in a co-processor board that provides the computational muscle to each PC server node.

In the initial phase of the LOFAR project only a single beam is formed from each receptor bandwidth slice and the first step in hierarchical beam forming could be executed at the receive cluster level, which reduces the effective output data rate by a factor five. Then five receive clusters and a set of beam forming processing nodes could be daisy chained by 10GbE links. With a proper forwarding process in each node a ring router is formed. One of the processing nodes is connected to a second ring, which combines the cluster set outputs and that also includes one or more nodes of the central processor. This set of beam forming processing nodes could be co-located at the central processing facility, or located at an individual station and then reduce the capacity requirement for the long-range network. The central processing cluster network could be organized as a three-dimensional torus [4] that supports an average inter node transport capacity of order Gb/s, which is adequate to support the cross correlation functionality. This situation is depicted in Fig.2, where 10GbE fibre optic links are used for each receive cluster of the inner array. The stations of the outer array have a single fibre optic link, requiring additional transceivers as repeaters for the remotest stations [5].

Instead of forming a limited set of beams in a hierarchical fashion for each station and cross correlating them, an all sky imaging mode could be provided that uses only the cluster signals of the inner array, together with the processing power available in all the nodes. However this beam forming operation would require twelve Tb/s throughput from the receive

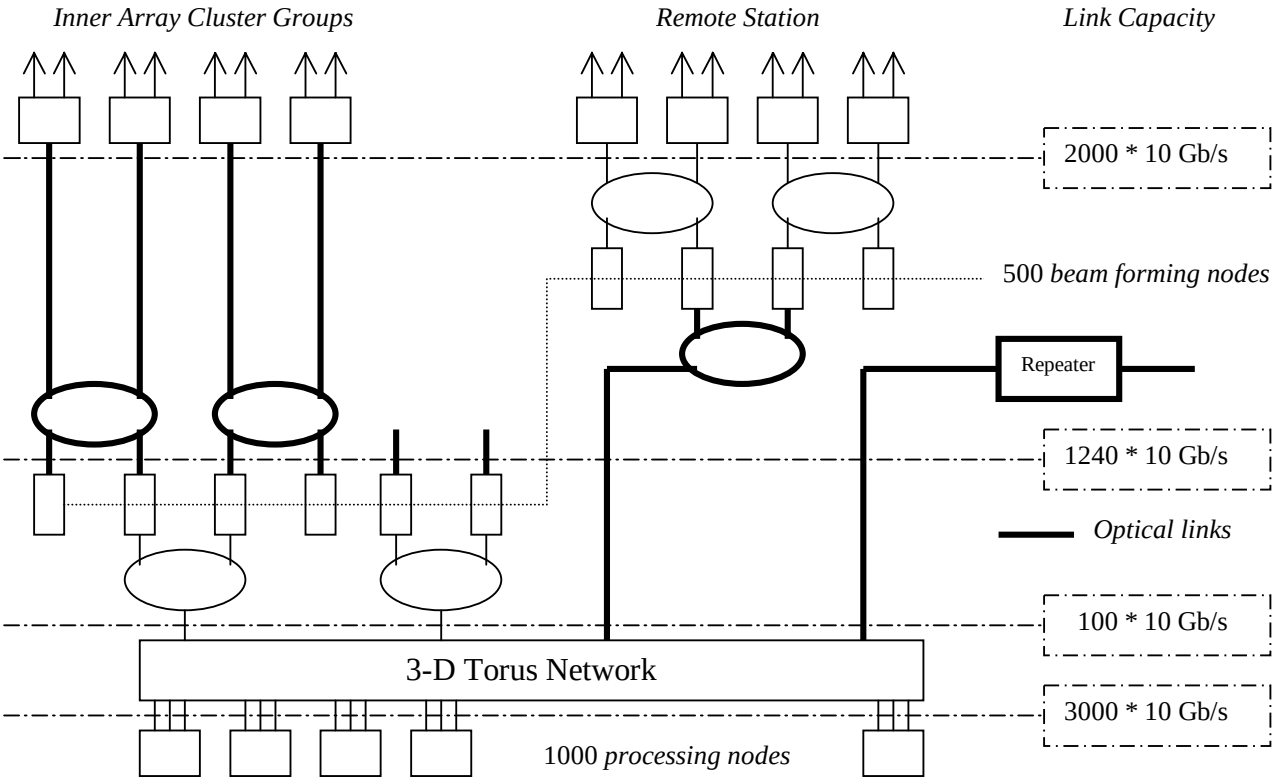

Fig. 2. Network architecture for the first phase of LOFAR based on 10GbE technology and ring routing

clusters to the two thousand processing nodes. This could be realized by replacing the three dimensional torus by a butterfly configuration of multi-port router devices, still using the same 10GbE interface cards. A final step is enhancement of the long-range network to support full sky imaging also by the remote stations. This could be realized in WDM / OTDM technology where many 10 Gb/s streams are multiplexed on a single fibre. This situation is visualized in Fig. 3.

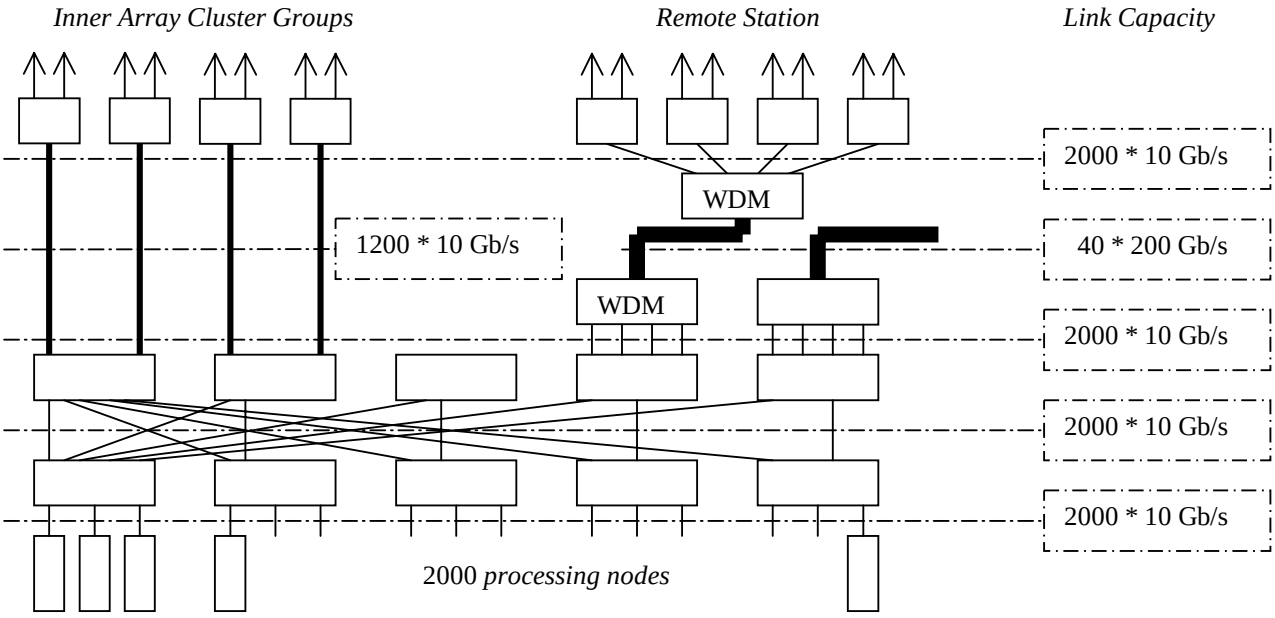

Fig. 3. Network architecture for the second phase of LOFAR including WDM/OTDM technology.

CONCLUSIONS

A network based on 10GbE transceiver and routing technology being developed for the emerging ten Gigabit Ethernet market is a viable option for LOFAR. We have shown that full 200 Gb/s data transport for 60 % of the stations out till 13 km from the centre is a system optimum after 2004 when the cost of trenching, fibres and transceivers is balanced. This gives complete freedom in locating hierarchical beam forming near the receive clusters or at the central processor.

The remote stations could operate at 10 Gb/s in the initial phase of the project requiring forms of data compression and reduction like local beam forming. When advanced WDM / OTDM data transport technology becomes available by 2007, it could be installed together with other planned processing hardware upgrades of the system and eliminates the need for station level beam forming. We postpone our most demanding applications till advancing technology provides high bandwidth processing according to Moore's law at the appropriate marginal system enhancement cost.

A network routing approach is proposed that allows all processing nodes in the network to be accessed efficiently by all receive clusters. This provides for dynamic reconfiguration of the processing constellation to meet observation requirements. A key feature is the bandwidth scalability, which allows network configurations to be simulated and evaluated using current one Gigabit Ethernet technology.

REFERENCES

[1] J.D. Bregman, "Concept Design for a Low Frequency Array," SPIE proceedings Volume 4015, March 2000

[2] J. Verhoosel, M. de Vos, E. J. van Veldhuizen, T. Koonen, H. de Waard, "A multi-terabit optical fibre network for a LOFAR telescope," Proceedings URSI GA, July 2002, ibid.

[3] K. v.d. Schaaf, "LOFAR Architectural Design Document," LOFAR-ASTRON-ADD-006, www.lofar.org

[4] C.M. de Vos, K. v.d. Schaaf , J.D. Bregman, "Cluster computers and Grid Processing in the first Radio Telescope of a New Generation," *Proceedings of the First IEEE/ACM International Symposium on Cluster Computing and the Grid*, Brisbane, May 2001, pp. 156-160

[5] H. Ou, M. v. Veelen, D. Kant, "Assessment of gigabit Ethernet technology for the LOFAR multi-terabit network," Proceedings URSI GA, July 2002, ibid