EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CERN/ECP 90-10
22 November 1990

CM-P00056661

Buses and Standards for LHC

H.Müller¹ J.Belleman¹ A.Bogaerts² D.Burckhart¹ R.Divìa¹
H.Döbeling⁴ C.Eck¹ G.Heyes¹ D.Linnhöfer² R.McLaren¹
C.Parkman¹ M.Pentney¹ J.Petersen¹ P.Ponting¹ J.F.Renardy³

CERN, Geneva, Switzerland

Abstract

The ECFA sub-working group on 'standards and buses' has investigated the role of both conventional and new buses for LHC. Three working teams present their results in specialized companion papers. This summary paper compares the use of Fastbus, SCI, VME-family buses and Futurebus+ using a common model of a LHC data acquisition system. The different buses were investigated for their capabilities to be used in the various areas of our standard model for an LHC system, and the results are shown in a survey table.

Both the model and the major conclusions are presented here. The most general conclusion is that buses will have an important future in LHC data acquisition experiments. The following paragraphs give an explanation.

Why standard buses

Though any model of LHC can only have significant uncertainties, several global problems can be identified for which the use of standards and buses provides a most reasonable solution: Massive connectivity of up to \(10^7\) channels, extremely large event sizes of \(10^7\) bytes/event and very high throughput of \(10^{10} - 10^{11}\) bytes/s. This will require:

- Uniform transmission properties of data paths
- No crosstalk and EMF sensitivity
- Unique addressing schemes for single and multiple access.
- Minimal, very well defined read-write protocols.
- Minimal, very well defined access protocols (arbitration) for fan in units in a data driven environment
- Error detection and recovery
- Live insertion and removal of parts
- A modular, mechanical framework which allows for replacement, renewal and maintenance.

¹CERN CH-121 Geneva 23
²Institut für Hochenergiephysik, ETH Zürich
³Saclay, France
⁴PSI, Villigen, Switzerland
• A very reliable power supply and cooling environment

All these points are obviously reasons why bus standards were created. It will require very innovative ideas and high investments into verification in order to replace the standard bus solutions by chip I/O specs, capton-foil systems, or meshes of private links. System-level specifications, available for most of the standard buses will need to be invented, and imposed for such systems.

Standards: An important feature of buses is the attribute standard which can be translated as: things fit together and have guaranteed and very well defined properties. It also means that industry will supply standard components, ready to fit into a system. In very large systems like LHC, 'standards' should therefore provide a safe environment for the chip- and board-level designer as well as for the system architect.

Are buses too slow? The role of standard buses for LHC is frequently seen together with performance in existing HEP experiments, in particular LEP. The overall performance in LEP’s data acquisition systems is much less than the bandwidth of the buses in use. An extrapolation of the use of standard buses to the very high bandwidth requirement in LHC seems therefore a priori excluded. It is however true that a “close to bus bandwidth” performance is a question of implementation. Particular choices of architectures and software as well as economy considerations strongly influence performance. With careful implementation choices, even conventional buses can be pushed to 40 - 100 Mbytes/s.

New, powerful bus generations are evolving, driven by interest from high end computing industry, to provide performance and interconnectivity much beyond the limits of conventional buses. Examples:

• VME64 practically doubles the performance of VMEbus
• Futurebus+ may achieve 3.2 Gbyte/s on 256 bit backplanes

• SCI, a point-to-point cable bus will have a performance of 1 Gbyte/s/node

Other buses like Fastbus could also be able to increase significantly their real performance if investments into better implementations are made.

2 How to use buses

Neither Fastbus, nor VME are used in HEP experiments at even half of their backplane bandwidth limit. The overall performance of large Fastbus LEP systems is even far from the bandwidth limit; though a performance of 150 Mbytes/s was achieved and published [7], LEP experts quote typically 10-20 Mbytes/s as peak performance! Responsible for the discrepancy are ‘implementation choices’ which either should be avoided for LHC, or the bandwidth requirement should generally be multiplied by a factor of 2 or 4.

Implementation choices: The ECFABUS team uses unscaled bandwidth figures and therefore assumes that implementation choices will be taken with highest priority for performance: A loss of performance can be due to:

• embedded ‘de-luxe’ software
• no parallelism for the sake of economy
• not making use of data driven concepts where applicable
• curing of ‘flaky’ implementations by slow speed operation
• general purpose designs with inferior performance than specific purpose
• module design with functionality specs rather than performance

Areas for new buses: The new bus standards stretch their bandwidth well into the 1 Gbyte/s range. Assuming we know how to make good use of it, the new buses promise to be excellent candidates for an LHC data acquisition system in the areas of the 3 main data streams:
- global trigger level 2 decision complex
- level 3 trigger processing farm
- data logger

These streams require bandwidth in the order of $10^8-10^9$ bytes/s (see Fig. 1) and are therefore problematic for conventional buses.

**Bus layers:** Buses will be specialised for certain layers of future systems, i.e. there will be 'no unique, single bus'. Special attention is required for the interfaces between different bus layers. *Conventional backplane buses* are more likely to be found around the local trigger level 2 area, while new buses are required for the global trigger 2, trigger 3 and event building stage.

**Conventional buses:** Buses like like Fastbus or VME are applicable wherever new buses are neither required nor economical. They are both an economical as well as a 'safe-ground' choice for the areas where data from many input channels is stored and locally processed. The max. bandwidth requirements here are typically 100 Mbytes/s.

**Backplane buses:** Buses like like Futurebus+ VME64 or Fastbus provide, via their short and well defined backplanes, very high speed local interconnections. This is the optimal environment for local processing and includes the physical framework for housing high density, power-hungry modules in a well defined, standard way. The interconnection between backplanes can be based on point-to-point buses or links, if the standard does not already provide it. Most backplane bus standards provide upgrade path from a conventional bus into a new bus.

**Point-to-Point buses:** The SCI Scalable Coherent Interconnect provides a novel way of interconnecting processor and memory nodes at very high speed which is largely independent of the number of nodes. Any node may send or receive data or commands to other nodes in the network. Topologies of interconnected ringlets can be easily changed by changing the cabling. Standard interconnections between commercial processors and backplane buses like VME are expected to become available soon together with the node chips for specific memory interface designs. SCI presents a very promising solution for the main data steams in LHC as described above.

**New features of new buses:** The availability of new features like caching, split transactions, virtual memory access will allow for better performance, more throughput and new data acquisition concepts. Further studies in connection with test-setups and system simulation are however required. Companion specifications like the CSR specification allow for merging of bus layers and permit to use the same software for different layers. CRC checking, available for SCI, will considerably improve error detection.

3 The ECFABUS model

The ECFABUS model (see Fig. 1) can only serve as a rather crude assumption. Its main purpose is to provide a common denominator for terms used within the ECFABUS teams. This model divides the LHC data acquisition into 3 crude areas:

- **detector level** including the trigger level 1
- **bus level,** situated further away from the detector receiving zero-suppressed data from detector channels after trigger 1 'Yes'
- **link level** where high bandwidth is required to link different parts of the detector for global triggers and event building.

**The Detector level:** Due to confinement in space and analog signals this part is unlikely to be implemented using buses. We have therefore excluded, with the exception of VX1bus the use of standard buses in this area.

**The Bus level:** Backplane buses are required in the local trigger-2 data concentrator
ECFABUS "MODEL" for BUS-LAYERS

Figure 1: The ECFABUS model

T1 : Trigger 1
LT2 : Local Trigger 2
GT2 : Global Trigger 2

ECFABUS, H.M.
9/10/96
units which in our model, receive digital de-randomized and zero suppressed data from the detector.

Units: ECFABUS uses the term unit for independent and uniform channel concentrators, receiving data after trigger level 1 'Yes'. They require internal buses in order to be able to act as data concentrators via embedded, fast hardware processors. Event-data is queue-stored in memories, directly accessible from both the local processors and from the output port of the unit. Each unit is independently concerned with local processing such as compaction and formatting. We assume that 1000 parallel units would be required for $10^6$ LHC channels however this number may be scaled.

The local trigger processing performed within the units is assumed to access only 1% of the data. At $10^8$ Hz input, the required unit internal bandwidth is $10^7$ B/s. Such requirements can be met either by conventional backplane buses (like Fastbus, or VME64) or new backplane buses. We conclude that units can be implemented using standard buses with the additional advantage of being economic and modular.

The Link level: This area is concerned with interconnection and routing of data. Since each unit has only very few (ideally one) output channel, the downstream part in our model consists of 1000 channels which need to be interconnected for the 3 data steams: global trigger 2, trigger 3 and data logger. In our model, this will require a total integrated bandwidth of $10^8$ bytes/s for each of them.

The global trigger 2: The input rate for this global trigger is equal to the unit's input rate of $10^5$ Hz since local processing cannot reduce the rate. Assuming a compression by 1/100 and access to only 1% of date from global trigger 2, the total bandwidth required for a global trigger 2 decision is estimated as $10^8$ bytes/s.

The event builder: Outputs from n units needs to be combined into one trigger 3 processor and, after a trigger 3 'Yes' into the data logger. Assuming a processor farm with n processors the eventbuilder is 1000 by n switching network. Such problems have undergone intensive SSC studies [6]. We are here only interested in an estimate of the required bandwidth. Bus-specific solutions are described in the companion papers [3] and [1].

In conventional event builders, full data is copied to the trigger processors, however only a 10% fraction may be used for the trigger 3 decision. We assume that data residing in the units is further compressed by a factor of 10. With a global trigger 2 decision filter factor of 1/100, the estimated bandwidth is $10^8$ bytes/s on the input of the trigger 3 processors.

The data logger: Assuming that trigger 3 reduces the input rate by 1/10, data is written at $10^2$ Hz with an event size of $10^6$ bytes/event. This requires a bandwidth of $10^8$ bytes/s and contributes to the load of the event builder network. It is unlikely that the data logging requires explicit use of buses. Commercial computers may be using buses like SCSI (or successors) for tape writing.

4 Applicability survey

A summary table (Fig. 2) has been established, showing our results on applicability of standard buses for the different areas of the LHC data acquisition model. The classifications are based both on the information presented in the companion papers [1] [3] [4] and the bus cross-reference paper [2] and are shown as 4 graduated steps:

- not applicable
- 0 possible
- + adequate
- ++ best choice

The best choice conclusions are presented in more detail in [5]
<table>
<thead>
<tr>
<th></th>
<th>FAST-BUS</th>
<th>FUTURE BUS+</th>
<th>SCI</th>
<th>VME</th>
<th>VXI</th>
<th>SCSI</th>
<th>MAX. CONSTRAINT</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>SPECIAL FAST ANALOG</td>
</tr>
<tr>
<td>T1 -&gt; T2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>(0)</td>
<td>-</td>
<td>10⁶-10⁷ Simple</td>
</tr>
<tr>
<td>BUFFERING</td>
<td>++</td>
<td>(+)</td>
<td>0</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>10-100 Ev. 10⁴-10⁵ Hz 1/100 Compr.</td>
</tr>
<tr>
<td>+ Loc. T2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LT2 -&gt; GT2</td>
<td>+</td>
<td>(+)</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10⁸ B/S from 1000 LT2 units</td>
</tr>
<tr>
<td>Buffer -&gt;</td>
<td>0</td>
<td>(+)</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10⁹ B/S from 1000 units</td>
</tr>
<tr>
<td>Ev. Builder</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Event -</td>
<td>0</td>
<td>+(0)**</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1000 x n -&gt; farm</td>
</tr>
<tr>
<td>Building</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ev. Build.</td>
<td>0</td>
<td>++</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10⁸ B/S</td>
</tr>
<tr>
<td>T3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D. Logger</td>
<td>+</td>
<td>++</td>
<td>++</td>
<td>0(?)</td>
<td>-</td>
<td>-</td>
<td>100 MB/S</td>
</tr>
<tr>
<td>-&gt; MASS -</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>++</td>
<td>4 x 10 MB/S</td>
</tr>
<tr>
<td>STORAGE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: Survey table ECFABUS

5 Acknowledgments

This work is a group effort of the ECFABUS team.
Special thanks also to: P. LeDu, J. R. Hansen, E. Kristiansen, D. Gustavson, C. Bee, H. Von der Schmitt, S. Cittolin and L. Mapelli as well as to all others who showed interest in our discussions.

References

[1] J.F. Renardy et al., SCI at LHC in these proceedings
[2] CSI buses and links group Buses Cross Reference version 1.0 ECFABUS team, edited by C. Parkman, CN division CERN
[3] D. Linnhöfer et al., Fastbus at LHC in these proceedings
[4] G. Heyes et al., VMEbuses and Futurebus+ at LHC in these proceedings
[5] L. Mapelli, Data Acquisition at LHC Plenary talk in these proceedings
[6] E. Barsotti et al., Effects of various event building techniques on Data Acquisition Architectures

Proceedings, Computing for High Intensity and High Luminosity Santa Fe 1990