The proposed Level-3 Trigger System for STAR

C. Adler [a], J. Berger [a], M. Demello [b], D. Flierl [a], J. Landgraf [c], J. S. Lange [a,1], M. J. LeVine [c], V. Lindenstruth [d], A. Ljubicic,Jr. [c], J. Nelson [e], D. Roehrich [f], E. Schäfer [g], J. J. Schambach [h], D. Schmischke [a], M. W. Schulz [c], R. Stock [a], C. Struck [a], P. Yepes [b]

[a]University of Frankfurt, August-Euler-Straße 6, D-60486 Frankfurt, Germany
[b]Rice University, Houston, Texas 77251, USA
[c]Brookhaven National Laboratory, Upton, New York 11973, USA
[d]University of Heidelberg, Philosophenweg 12, 69120 Heidelberg, Germany
[e]University of Birmingham, Birmingham B15 2TT, United Kingdom
[f]University of Bergen, Allegaten 55, 5007 Bergen, Norway
[g]Max-Planck-Institut für Physik, Foehringer Ring 6, 80805 München, Germany
[h]University of Texas, Austin, Texas 78712, USA

Abstract

The level-3 trigger system of the STAR experiment will in the final stage consist of a farm of 24 ALPHA/Linux processors, interconnected by SCI (Scalable Coherent Interface). The system will perform online tracking of $N_{track}\geq$8000 tracks per event ($N_{point}\leq$45 per track). The track data will be transfered to a global level-3 CPU (expected data transfer rate $\simeq$48 MB/s), performing online event analysis tasks (e.g. invariant mass reconstruction) with a design trigger input rate of R=100 Hz (R=20 Hz for a prototype system).

I. Introduction

The RHIC accelerator at Brookhaven National Laboratory, USA, will start in 1999 to investigate *Au+Au* collisions with $\sqrt{s}\leq$200 A·GeV and *p+p* collisions with $\sqrt{s}\leq$500 GeV.

The STAR experiment [1] is a large scale, cylindrical, symmetric 4π-detector at one of two main RHIC interaction points. Data taking will start in 1999 with a full size TPC (Time Projection Chamber, R_{in}=0.6 m, R_{out}=2 m) with 24 TPC sectors, 6912 pads each. TPCs are specificaly suitable for detecting high density charged particle fluxes in high multiplicity nucleus-nucleus events.

II. The STAR Trigger Architecture

The STAR trigger system is subdivided into 4 hierarchic levels. The level-0 input rate is 10^5 Hz, the first three levels reduce the rate by one order of magnitude each. The level-3 trigger is supposed to reduce an input rate of 10^2 Hz to the final DAQ rate of R_{tape}=1 Hz at an expected TPC event size of $\simeq$15 MB.

A task example for the level-0 trigger is the selection of central and peripheral *Au+Au* events based upon multiplicity (function of impact parameter). A task example for a combined level-1/-2 trigger is selection of events with a vertex near the beam crossing point (using information of a vertex position detector).

The tasks of the level-3 trigger are event selections based upon the online reconstructed track parameters of each particle. Two examples are:

- *for Au+Au collisions:* online invariant mass reconstruction of $J/\psi\rightarrow e^+e^-$, as suppression of J/ψ production is commonly regarded as one the the most promising signatures of the quark-gluon plasma, and
- *for p+p collisions:* filtering of 700 pile-up events in the TPC per one level-0 trigger.

Other applications as beam background rejection and online jet finding have been proposed, too.

III. The STAR DAQ architecture

The level-3 trigger design is embedded into the STAR DAQ system [2], working closely to all DAQ components.

Each physical TPC sector is mapped onto one VME crate, containing a *Sector Broker*, i.e. Motorola MVME-2306 VME board, carrying a PowerPC 604 (300 MHz, VxWorks), as the TPC sector master controller. The Sector Broker carries two VMETRO [3] C2D PMC-SCI interfaces for *(a)* raw data transfer to the main STAR event builder and *(b)* connection to the level-3 track finder CPU.

Moreover, each DAQ crate also contains six VME receiver boards, each carrying three mezzanine cards with

- one Intel i960 CPUs (33 MHz, VxWorks) for *(a)* data formatting, *(b)* initiating the VME raw data transfer and *(c)* running the level-3 cluster finder,
- 4 MB of dual-ported VRAM for buffering and pipelining of raw data of 12 events.

IV. Level-3 Trigger Architecture

The level-3 trigger scheme consists of two main parts:

- The *sector level-3* part (shown in Fig. 1) is mapped onto one physical TPC sector. It contains *(a)* the level-3 cluster finder (described in Section V) and *(b)* the level-3 track finder (described in Section VI). Data transfer of cluster

[1]Corresponding Author, Email soeren@bnl.gov

data and track data is performed by SCI (described in Section VIII).

- The *global level-3* part (shown in Fig. 2) consists of one master CPU for the whole STAR TPC, collecting all track data via SCI and issuing the level-3 decision.

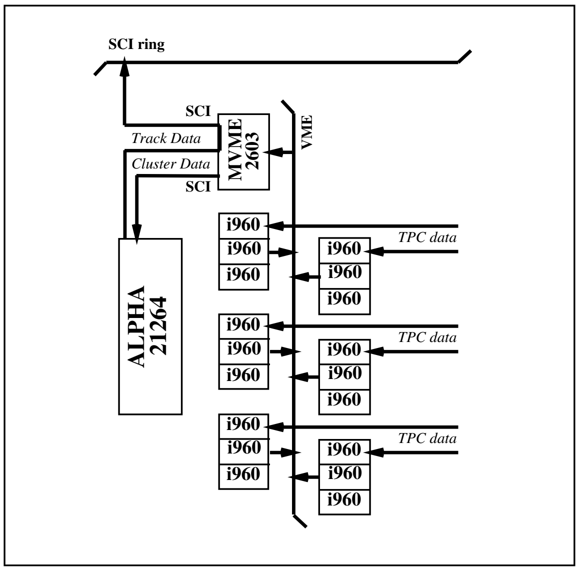

Figure 1: Level-3 Sector Architecture.

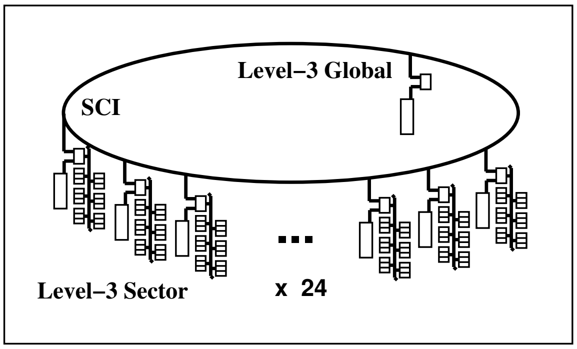

Figure 2: Level-3 Global Architecture.

The development of the level-3 trigger can be subdivided into two main stages.

- In the first stage, envisaged for 1999, eight TPC sectors in a $\Delta\varphi \simeq 180^o$ topology are planned to be connected. The level-3 trigger will employ TPC data only, and the input trigger rate is estimated to be R=20 Hz.

- In the second stage, envisaged for 2000, all 24 TPC will be connected. The level-3 trigger will employ additional information of SVT (Silicon Vertex Tracker) and EMC (Electromagnetic Calorimeter), and the final design value for the input trigger rate is R=100 Hz.

As will be shown in Ch. VI, currently available CPUs are not capable to perform high multiplicity tracking within the given time constraint. Thus, in the second stage probably more than one CPU per TPC sector must be used, implying programming parallelization techniques.

V. THE CLUSTER FINDER

For a TPC readout, one ADC channel is indexed by a pad number ($r\varphi$-direction, e.g. 88 pads for the most inner padrow at R_{in}=0.6 m) and a drift time bin number (z-direction, 512 time bins per pad). Clusters are continous ($r\varphi$,z) regions with ADC>ADC-threshold.

In a first step, for each TPC cluster the center-of-gravity (weighted mean according to ADC values) is calculated to obtain particle hit xyz-coordinates.

The cluster finder algorithm runs on the Intel i960 CPUs, implemented on the DAQ receiver boards. The number of i960s is 18 per TPC sector, 432 for the whole TPC. With their DAQ tasks (cf. Section III) the i960s are not completely occupied, thus additional execution of the level-3 cluster finder is not CPU time critical.

Input to the cluster finder are zero-suppressed TPC raw data, stored in the VRAM. The output cluster data, i.e. *(a)* cluster center-of-gravity and *(b)* cluster total charge (ADC sum), are sent via VME to the Sector Broker, which itself ships the data via SCI to the level-3 track finder CPU (expected data transfer rate of $\simeq$3 MB/s per TPC sector).

The time constraint is $\tau_{cluster} \leq$10 ms (input rate 100 Hz). Benchmarks on the i960 were performed for 600 clusters (realistic *Au+Au* scenario) on the TPC's most inner padrow. The position resolution (reconstructed minus Monte-Carlo generated cluster position) of $\Delta(r\phi) \simeq 37\mu$m and $\Delta z \simeq 13\mu$m could be obtained with an algorithm within $\tau_{cluster}$=7.5 ms. The clusters and reconstructed centers-of-gravity are shown in Fig. 3. If two clusters are merged, an additional deconvolution subroutine must be started, consuming 6.0 % more CPU time than in case of two separated clusters.

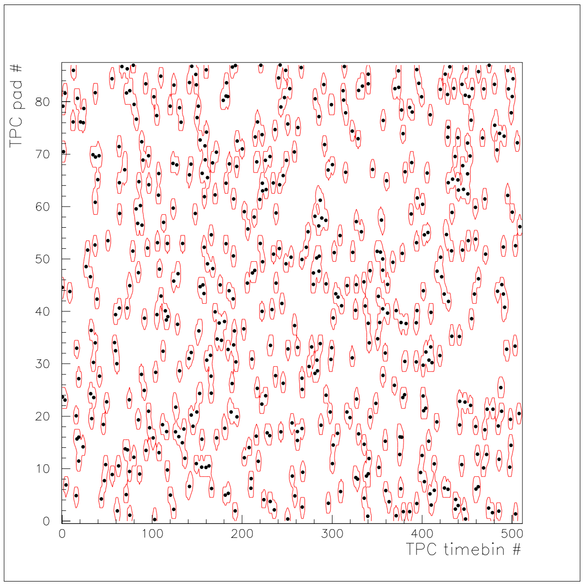

Figure 3: Level-3 cluster finder example (600 clusters on the most inner TPC padrow). The black spots indicate the centers-of-gravity, found by the cluster finder algorithm on the Intel i960 within $\tau_{cluster}$=7.5 ms.

VI. THE TRACK FINDER

In case of an *Au+Au* collision, the track finder algorithm must be able to fit at least $N \simeq 400$ tracks per event per TPC sector, each consisiting of $N_{point} \leq 45$ points (given by the number of padrows).

The fast track finder algorithm has specificaly been developed for the level-3 trigger project [4]. It employs *conformal mapping* (transformation of a circle[2] into a straight line), followed by a fit with a *follow-your-nose* method.

Input to the track finder are the cluster data, transfered via SCI from the Sector Broker. The dispatch of *(a)* the 12 pipelined events in the VRAM, and *(b)* 18 parts of cluster data (one per i960) is performed by a trigger *token* scheme, i.e. a handshaking protocol between the Sector Broker and the track finder CPU. Output track data are transfered by a 2-step SCI transfer: *(a)* from the track finder CPU to the Sector Broker, *(b)* via the SCI DAQ ring to the global level-3 CPU.

The track finder time constraint of $\tau_{track} \simeq 110$ ms is given by the buffer time of 12 pipelined events (12×10 ms), minus the time being necessary for cluster finding $\tau_{cluster} \leq 10$ ms (cf. Section V). The track finder code was benchmarked on several Linux CPUs; the results are given in Fig. 4.

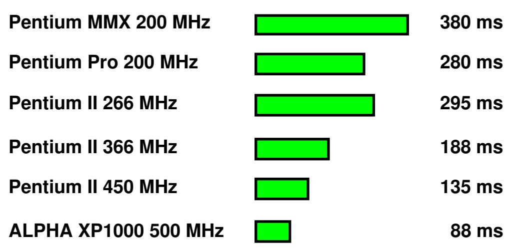

Figure 4: Level-3 track finder benchmark. CPU time τ_{track} for different CPUs (Linux Kernel 2.0.xx or 2.2.x), egcs 1.0.3 (optimizing flag -O2) for a simulated *Au+Au* event with 400 tracks per TPC sector.

According to the benchmark results, the ALPHA XP1000 500 MHz workstation was the only CPU to achieve the time requirement, if one restricts the number of CPUs per TPC sector to one. Based on these results, the ALPHA 21264 was chosen for the first level-3 track finder implementation. An example of 1000 π^+/π^- particle tracks is shown in Fig. 5. Other architectures (e.g. Pentium III 733 MHz) are candidates for future level-3 extensions.

VII. ALPHA

An ALPHA XP1000 500 MHz (128 MB RAM) was used as a prototype track finder CPU. It contains an ALPHA 21264 chip [5], a RISC CPU with only 160 instructions.

For reasons of compatibility to the STAR offline software analysis framework, Linux was chosen as operating system.

[2] In the STAR solenoid magnetic field of B=0.5 T charged particle tracks can be parametrised as helices, being visible as circles in an xy-projection.

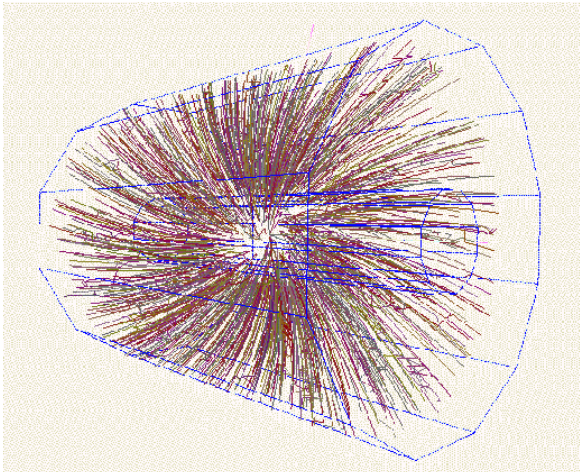

Figure 5: Level-3 track finder example. 1000 π^+/π^- particle tracks in the STAR TPC are shown. Tracking was performed on an ALPHA XP1000 within τ_{track}=9.6 ms per TPC sector.

Linux kernel 2.2.3 is running stably. As the MILO boot loader was not available for the XP1000, booting could only be achieved by switching the machine into the SRM console mode before Linux boot.

The ability to execute four floating point instructions per cycle leads to a very high floating point performance, e.g. faster than a Pentium II (clockcycle scaled) by a factor ∼6.5 for a looped floating point statement like `f=f/2.0+f/3.0`. However, an integer division hardware instruction is not available and should be avoided in programming.

Concerning level-3 specific track finder requirements (memory access to cluster data), two ALPHA features are important: *(a)* internal 4MB L2 cache (Pentium only external L2 cache) and *(b)* memory bus width of 128 bit (Pentium 64 bit), memory access automatically scheduled by TSUNAMI-D chip [5].

The ALPHA chip is statically scheduled, i.e. the performance depends on the sequence of instructions. Specific load/store-sequence optimizing techniques (”*lending the compiler a hand*”) can increase the performance significantly.
Example code A:

```
dst[i]=func(src[i]);
```

Example code B:

```
tmp=src[i]; tmp=func(tmp); dst[i]=tmp;
```

The CPU time fraction of A/B shows the completely different behaviour of the different platforms: 1.00/1.15 in case of ALPHA and 1.00/0.66 in case of Pentium.

The ALPHA 21264 is the first ALPHA chip with an implemented hardware SQRTF intruction, but not supported by egcs 1.0.3 (Redhat 5.2 Linux installation). The usage of a specific COMPAQ Portable Math Library [6] improved the track finder performance ($\simeq$30 sqrt() calls per track) by $\Delta\tau/\tau$=38 %.

to the track finder time of $\tau_{track}\leq$110 ms, the data transfer time is negligible. Moreover, the participating CPUs (i960, ALPHA 21264) are low-endian (byte order 1-2-3-4), thus additional byte order conversion before or after data transfer is not necessary. However, the level-3 track data for all 24 TPC sectors increase the load on the SCI DAQ ring by additional 48 MB/s, corresponding to $\simeq$30 % of the TPC raw data bandwidth.

VIII. SCI

The SCI (Scalable Coherent Interface) [7] bidirectional interface standard (A64/D16) utilizes transfer of small data payload of 16 or 64 bytes at a high clock frequency of 200 MHz. The maximum bandwidth is 200 MB/s, signals are transmitted via 18 pin differential ECL. The low latency of 2.3 μs (64 byte transfer) is achieved by *(a)* point-to-point connections (eliminating "starvation") and *(b)* "RISC like" protocols (reducing overhead).

The level-3 trigger system uses D310 PCI-SCI interfaces by DOLPHIN^{tm} [8] . Prototype measurements of SCI bandwidth and latency as a function of DMA buffer size (track data) are shown in Fig. 6 (results compatible with former studies [9]).

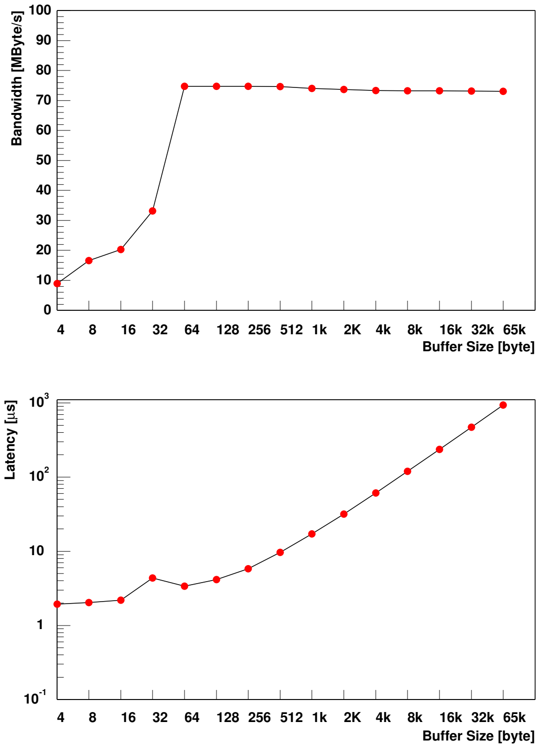

Figure 6: SCI bandwidth (top) and SCI latency (bottom) as function of DMA buffer size (track data) for two point-to-point DOLPHIN D310 PCI-SCI adapters.

As an SCI driver for ALPHA 64 bit PCI bus interface was not available (different handling of SCI address translation tables), the development of a low level SCI Linux device driver was launched within the level-3 project. The SCI programming principle of *remote memory access* differs from *sending packet* strategies as e.g. Gigabit Ethernet. Moreover, the background DMA (Direct Memory Access) capability leaves the CPU completely free for other tasks (e.g. track finding).

SCI was chosen in order to keep the transfer time within the level-3 system as short as possible. Assuming a point-to-point SCI bandwidth of B=72 MB/s (prototype result, cf. Fig. 6 top, DMA buffer size >64 kB), the cluster data transfer (from i960 to ALPHA) of 30 kB per event takes $t\simeq$0.4 ms and the track data transfer (from sector level-3 CPU to global level-3 CPU) of 20 kB per event takes $t\simeq$0.3 ms. Compared to the track finder time of $\tau_{track}\leq$110 ms, the data transfer time is negligible.

IX. GLOBAL LEVEL-3 TRIGGER

One *global* level-3 CPU receives the track data from all sector level-3 CPUs. It is connected to the SCI DAQ ring via a Motorola MVME 2306, in this case not serving as a Sector Broker, but only as interrupt handler. The SCI data are sent directly to the global CPU, the SCI messages (end-of-DMA, generating an interrupt) are only sent to the MVME.

The global level-3 CPU performs *(a)* track merging for tracks of different sectors, *(b)* a level-3 decision algorithm based on the *all* track data (e.g. invariant mass reconstruction) and *(c)* issues the level-3 yes/no decision as SCI message.

The total track data transfer rate is expected to be as high as 48 MB/s, therefore a fast and efficient memory management is mandatory. More specificaly large pieces of physical memory must be locked (until release by a token manager message) in order to avoid memory paging.

X. SUMMARY

The STAR level-3 trigger is planned to be a SCI interconnected ALPHA processor farm, performing online tracking of $N_{track}\geq$8000 particles with a design input rate of R=100 Hz. The components (cluster finder, track finder, SCI) have been benchmarked, a large scale prototype system (1/3 of the final design, R=20 Hz) is envisaged for STAR data taking in 11/99.

XI. REFERENCES

[1] STAR collaboration, "STAR Conceptual Design Report", Lawrence Berkeley Laboratory, University of California, *PUB-5347*, June 1992.
[2] A. Ljubicic, Jr. et al., "Design and Implementation of the STAR Experiment's DAQ", *10th IEEE Real Time Conference*, 9/22-26/97, Beaune, France.
[3] http://www.vmetro.com
[4] P. Yepes, "A Fast Track Pattern Recognition", *NIM A* 380 (1996) 582.
[5] http://ftp.digital.com/pub/Digital/info/semiconductor/
[6] http://www.unix.digital.com/linux/cpml.htm
[7] SCI (Scalable Coherent Interface) Standard Compliant, *ANSI/IEEE 1596-1992*.
[8] http://www.dolphinics.no
[9] J. J. Schambach et al., "Performance tests of the DOLPHIN and VMETRO PCI-SCI-Bridges", *CHEP 98, Computing in High-Energy Physics*, 8/31-9/4/98, Chigaco, USA.