

L2 Data Movement

Robert D. Martin

University of Illinois at Chicago

6 February 1999

L2 Bandwidth and budget/event @ 10KHz

<u>Link</u>	<u>MB/s</u>	<u>KB/event</u>	
G-Link	106	10.6	
Cypress Hotlink	16	1.6	
Mbus DMA	~80	8	320MB/s nominal
Mbus Prog	~20	2	.5-1 μ sec/16B
VME Block	>10	1	
VME Prog	~1	.1	.5-1 μ sec/4B

% Capacity used

Crate	Worst In /1.6KB	Mbus In/ 8KB	Mbus Out/2KB	Worst Out/1.6 KB	L3 /1KB (10%)	
Cal	22 % fixed	40% fixed	12%	5%	2%	
CTT	13% max	16% max	30%	19%	6%	Limit # tracks out
PS	12% max	26% max	6%	2%	1%	Avg < .3*max
MU to SLIC	6%	X	X	1%	X	
Slics to Mu Alpha	1%	10%	2%	3%	1%	
Global	19%	13%	X	X	3%	

Worst Case Latencies

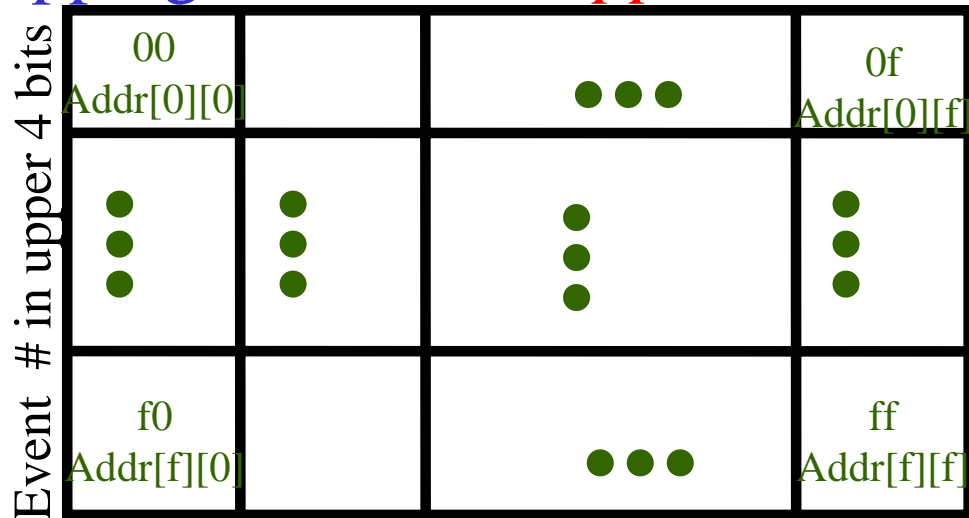
Wait = mean time others have to wait to arbitrate for the bus

$W = P(\text{busy}) * (T/2) * (1/N)$; N = Number of sources

System	Latency	Wait
Mbus DMA: L2Cal	38 μs	0.8 μs
Mbus Out: L2CTT	30 μs	2.3 μs
		(assume N=2)
L3: L2Global	30 μs	0.45 μs
L3: L2CTT	60 μs	1.8 μs

New Event Arrival

- MBus broadcast addresses mapped to 16 events with up to 16 sources per event
 - $mb_ad\langle 7:4 \rangle = \text{Event number}$
 - $mb_ad\langle 3:0 \rangle = \text{Source number}$
- This mapping stored in *Mapper* in DMA engine



New Event Arrival

- MBT broadcasts data with encode broadcast addr
- Every L2Alpha receives data in FIFOs and transfers to PCI address specified in the Mapper
- Interrupt fires when FIFOs on **all** L2Alphas have been drained
- Addresses for (N+16)th event placed in Mapper on all L2Alphas
- Last address written stored on Administrator for error detection

Preprocessor Output

- When preprocessor Worker nodes finish processing event, output needs to be sent to L2Global
- Transfer done using MBus Programmed I/O (PIO)
- Output data written to MBus address for appropriate output port on MBT.
- PIO transfer charged to Preprocessor time budget; no FIFOs decoupling L2Alpha from Mbus transfer

Interprocessor Communication

- MBus PIO used for Administrator-Worker and Administrator-MBT communication
- Administrator and Worker communicate via mailboxes in MBus space mapped to locations in each L2Alpha's main memory
- Time critical VME communication should be avoided due to latency introduced by VBD transfers

Interrupts

- Mbus
 - Only new event arrival
- VME
 - SCL_INITIALIZE: MBT to Admin
 - Worker reset: Admin to Workers
 - TCC
- Hard reset of L2Alphas via VME (not really an interrupt)

Modes of Operation

Two modes of operation in Multi-Worker crate

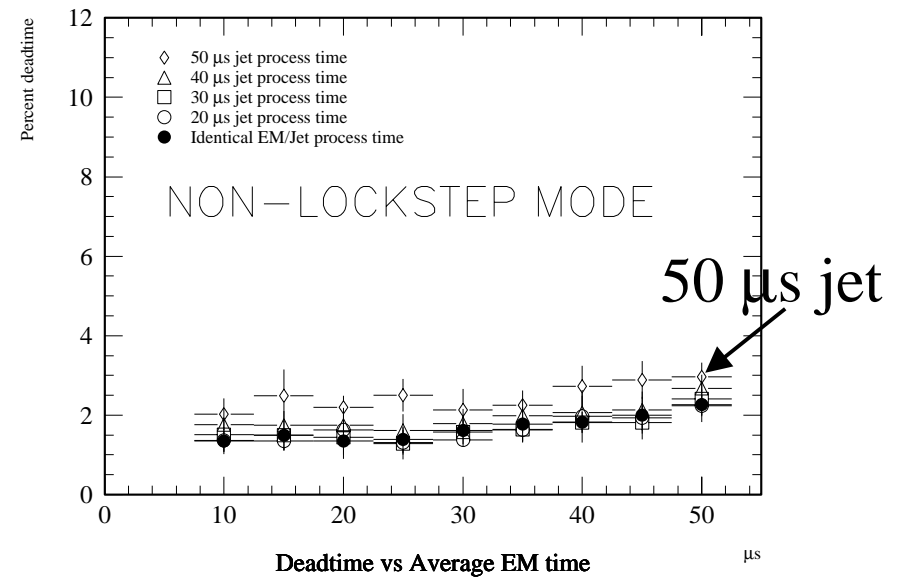
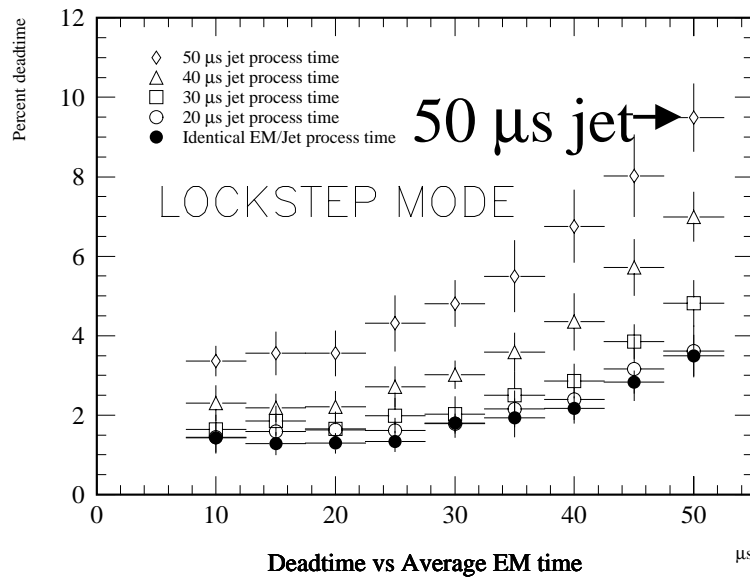
- Worker synchronous (lockstep)
 - Preprocessors
 - All Workers on same event
 - Must wait for slowest Worker
 - L2Global
 - “ping-pong”
- Worker asynchronous (non-lockstep)
 - PreProc: Worker moves on after processing evt
 - L2G: Worker moves to next unprocessed event

Queuing Simulations

- Use RESQ package from IBM
- Baseline system meets system requirements if:
 - Median Preprocessor time of roughly 50 μ sec
 - Median Global time of roughly 50 μ sec
 - Buffers placed between all elements
- Concerns:
 - Correlation between Cal Workers
 - Retreat paths for Cal and Global

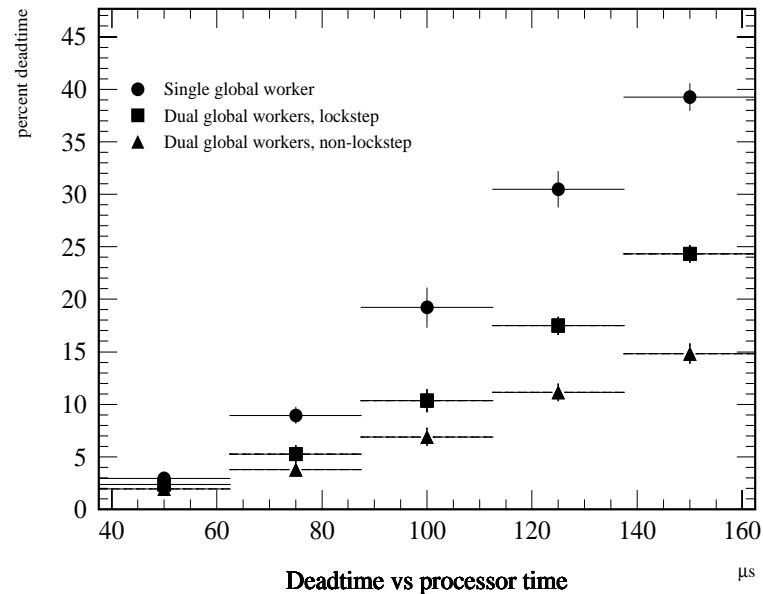
Calorimeter Preprocessor

- Times
 - E_T : (nearly) constant time 45 μsec
 - EM, jet: vary between 20 and 50 μsec



L2 Global Retreat Path?

- Additional workers
 - event synchronous mode (lockstep)
 - event asynchronous mode (non-lockstep)



Simulation Conclusions

- L2Cal
 - Deadtime from worst-of-n situation
 - Hope for lockstep; prepare for non-lockstep
- Global
 - Minimal gain from 2 processors in lockstep mode
 - More of a gain from non-lockstep mode with Administrator reserializing the results
 - Only way to get linear decrease in deadtime is to decrease input rate