BaBar Online Dataflow

Chris O'Grady, Amedeo Perazzo, Matt Weaver

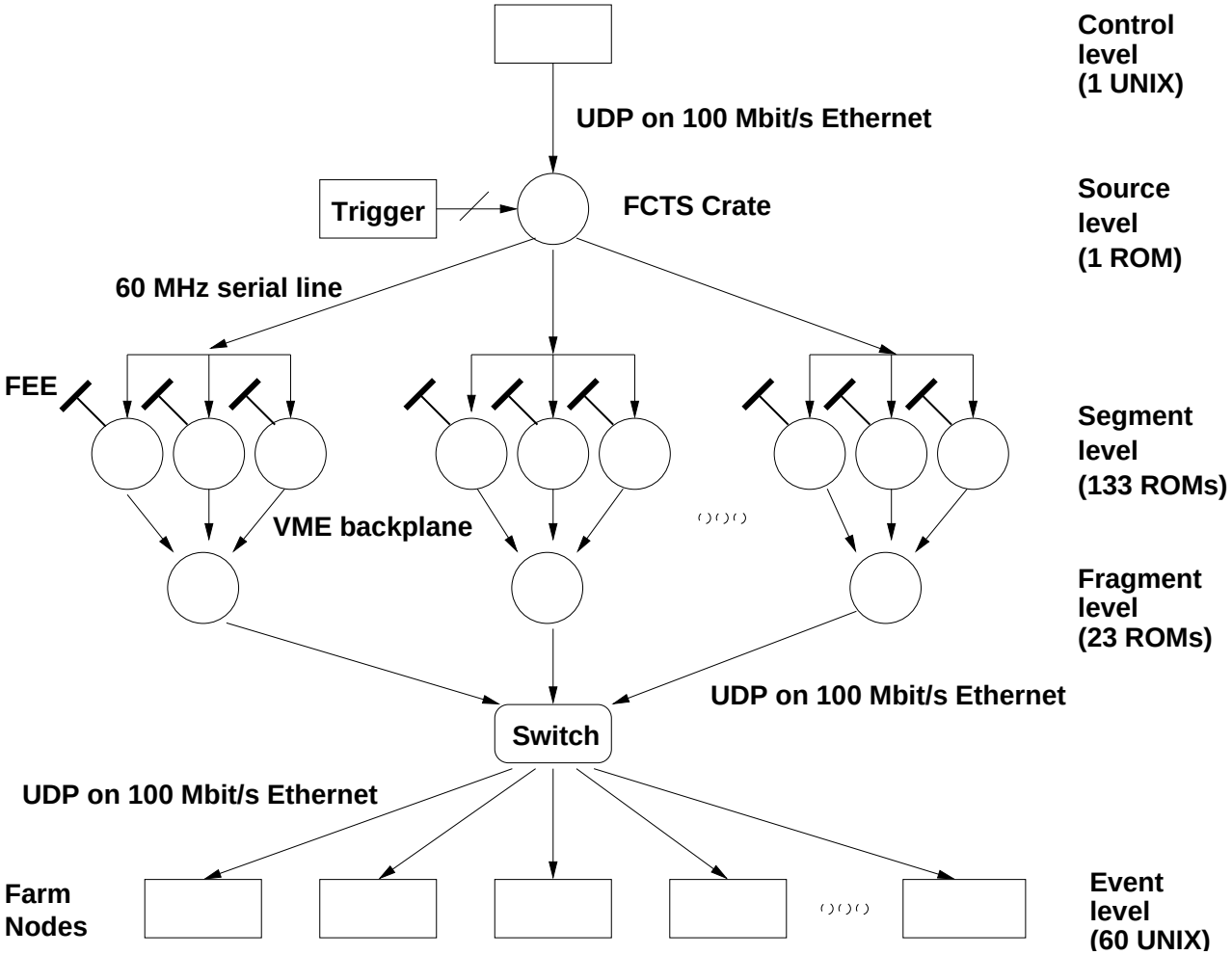
February 16th, 2002

1. Introduction to the ODF System
2. System Monitor
3. ODF Model
4. Limits of the Model
5. Bottlenecks
6. FEE Buffer Model
7. Upgrade Plan

BaBar Belle Workshop, TRIUMF

Introduction to the ODF System

- ODF handles DAQ and processing from BaBar Front End Electronics
- Delivers complete events to Level 3 Software Trigger
- L3 is a farm of 60 UNIX workstations
- 157 ReadOut Modules in the system located in 24 VME crates
- Network: switched 100 Mbit/s Ethernet



- ODF able to receive 1 Mbyte input data @ L1 trigger rate of 4 kHz, filter it to about 24 kbytes of event data

System Monitor

- To understand ODF system a monitor application (VMON) has been developed.
- Some of the quantities monitored by VMON include:
  - detector data sizes;
  - processing speed of feature extraction (FEX) code;
  - size of FEX output data;
  - VME DMA rates;
  - other event building performance parameters;
  - processing speed of the Level 3 trigger;
  - errors generated by the system.
- A model to describe the ODF performance has been developed using VMON together with few other bench tests.
- This model assign a processing time to all the ODF components.
- The system is able to operate at the frequency which corresponds to the slowest of the processing times before asserting back pressure and, hence, causing dead time.
- In order to extrapolate to future detector conditions we assume that the event size will scale with the sum of the beam currents. The extrapolation is done by comparing a cosmic run (i.e. no beams) with a normal data taking run.

ODF Model (1)

In 2007 we expect I_{HER} ≈4000 mA and I_{LER} ≈2000 mA

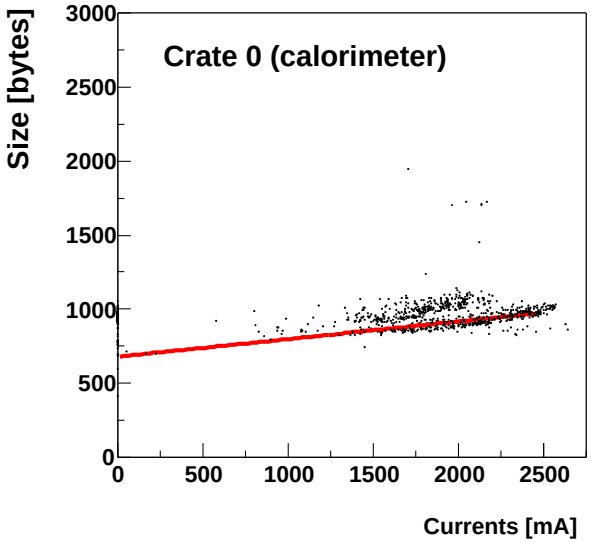
Time to process an event in microseconds (batch 2) (I+: 4000mA I-: 2000mA)

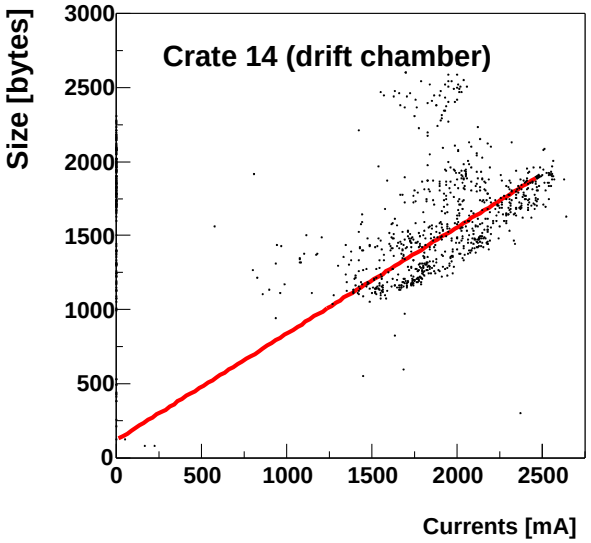
SYS (FESZ /FEXSZ)	GLNK	I960	PCIN	MPCN	CPUN	VME	PCI0	MPC0	CPU0	NET	SWTC
SVT (1096/ 1113)	75	62	13	31	95						
SVT						89	17	89	254	223	
DCH (8636/ 2213)	239	125	75	162	275						
DCH						67	17	89	252	221	
DRC (2463/ 1232)	20	102	24	53	158						
DRC						99	19	99	272	246	
EMC (9600/ 139)	0	105	75	152	215						
EMC						139	10	56	201	139	
EMC (6500/ 95)	0	81	51	103	144						
EMC						95	7	38	168	95	
IFR (3800/ 814)	19	104	33	69	52						
IFR (2700/ 446)	19	96	23	48	52						
IFR (3300/ 461)	19	100	27	57	52						
IFR						87	16	87	252	217	
EMT (5900/ 1503)	80	91	51	110	172						
EMT						60	11	60	199	150	
DCT (1226/ 336)	19	55	11	23	30						
DCT (502/ 688)	16	62	6	16	97						
DCT (391/ 282)	13	44	4	10	50						
DCT						73	10	52	186	131	
GLT (610/ 830)	73	31	8	20	20						
GLT						41	6	33	148	83	
BBR(983803/50629)											25

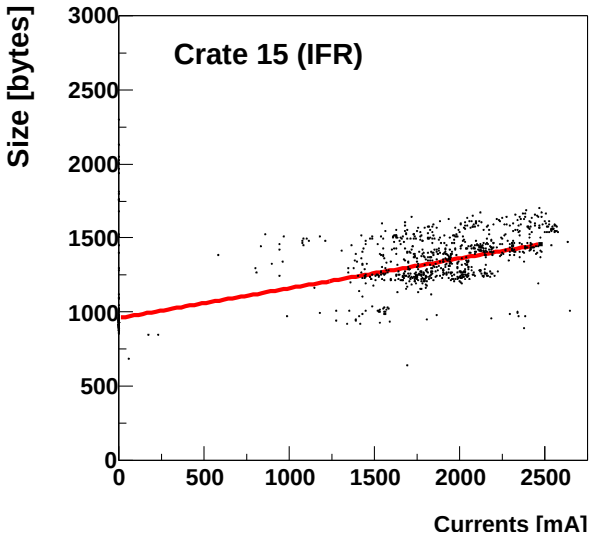
GLNK: transport on fron-ent optical fiber link to intermediate store
I960: transport from intermediate store to PPC memory
PCIN/0: PCI bus usage on slot0/slotN ROMs
MPCN/0: MPC bus usage on slot0/slotN ROMs
CPUN/0: CPU usage on slot0/slotN ROMs
NET: transport on 100 Mbit network

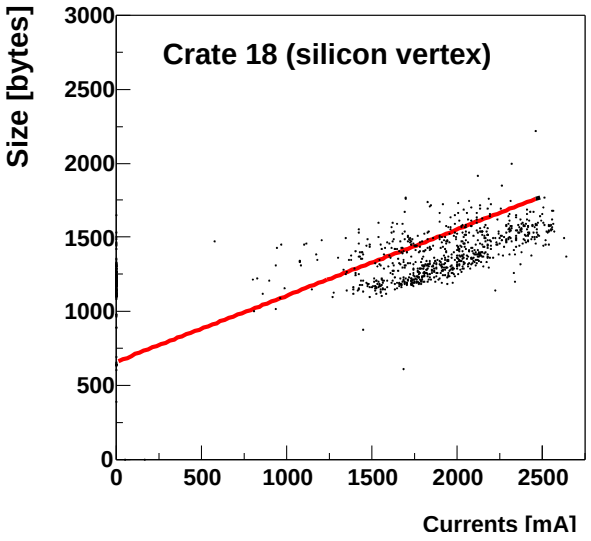
ODF Model (2)

- We tried to verify the event size scaling hypothesis with the event size extracted from various runs taken over the last two years



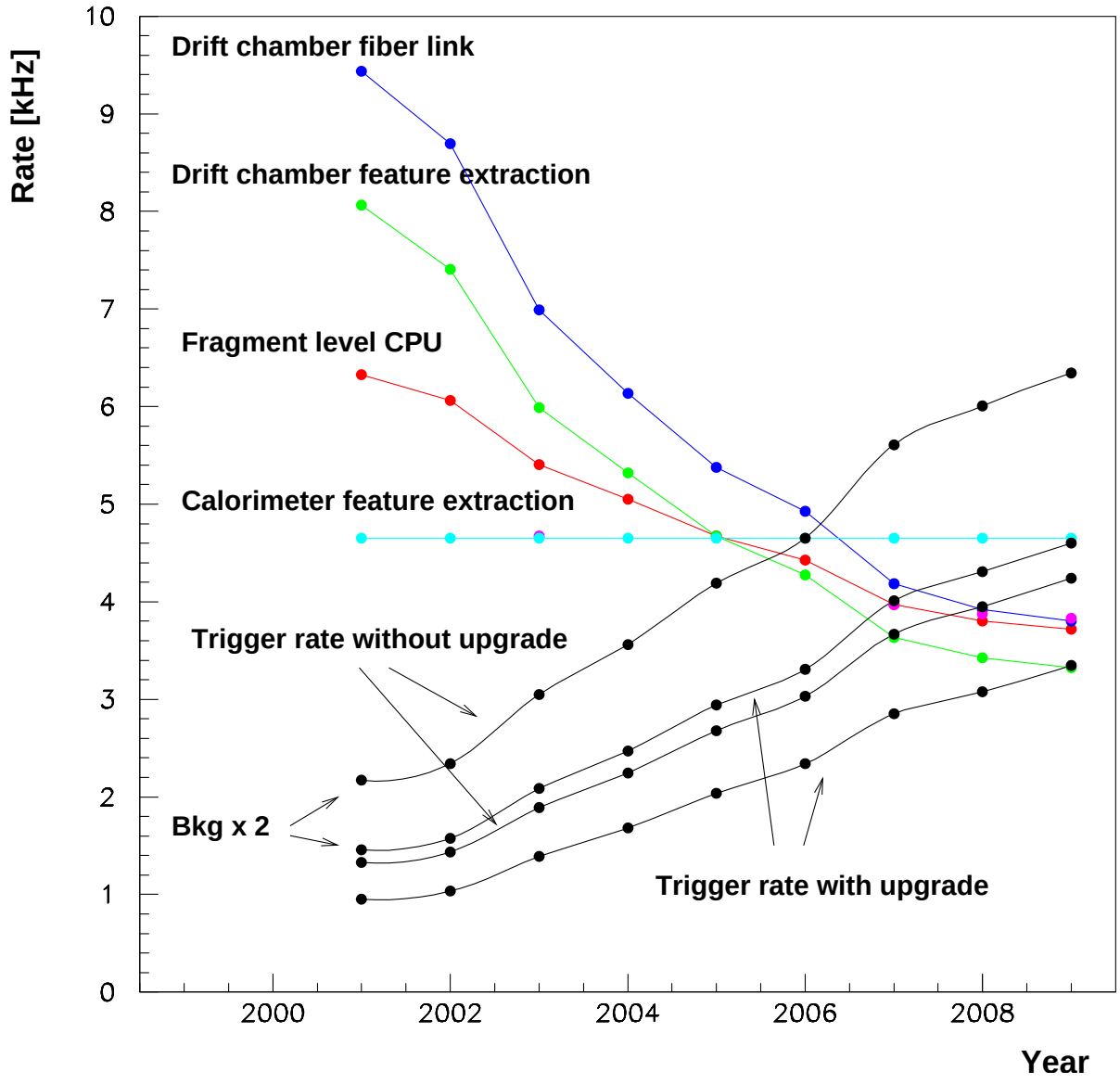






ODF Model (3)

- We compared our projections with the trigger rate estimated by the trigger group.
- Trigger rates consider four cases: upgrade/non-upgrade, expected bkg and bkg x2.



Limits of the Model

- Assumption event size scales with sum of the currents could be naive;
- model doesn't describe the buffering in FEE;
- we could get bad non linear behavior if we get close to the edge (e.g. bus thrashing on the PCI or the VME bus);
- some detectors could be unusable at the occupancies implied by these event sizes (e.g. first super-layer in the drift chamber).

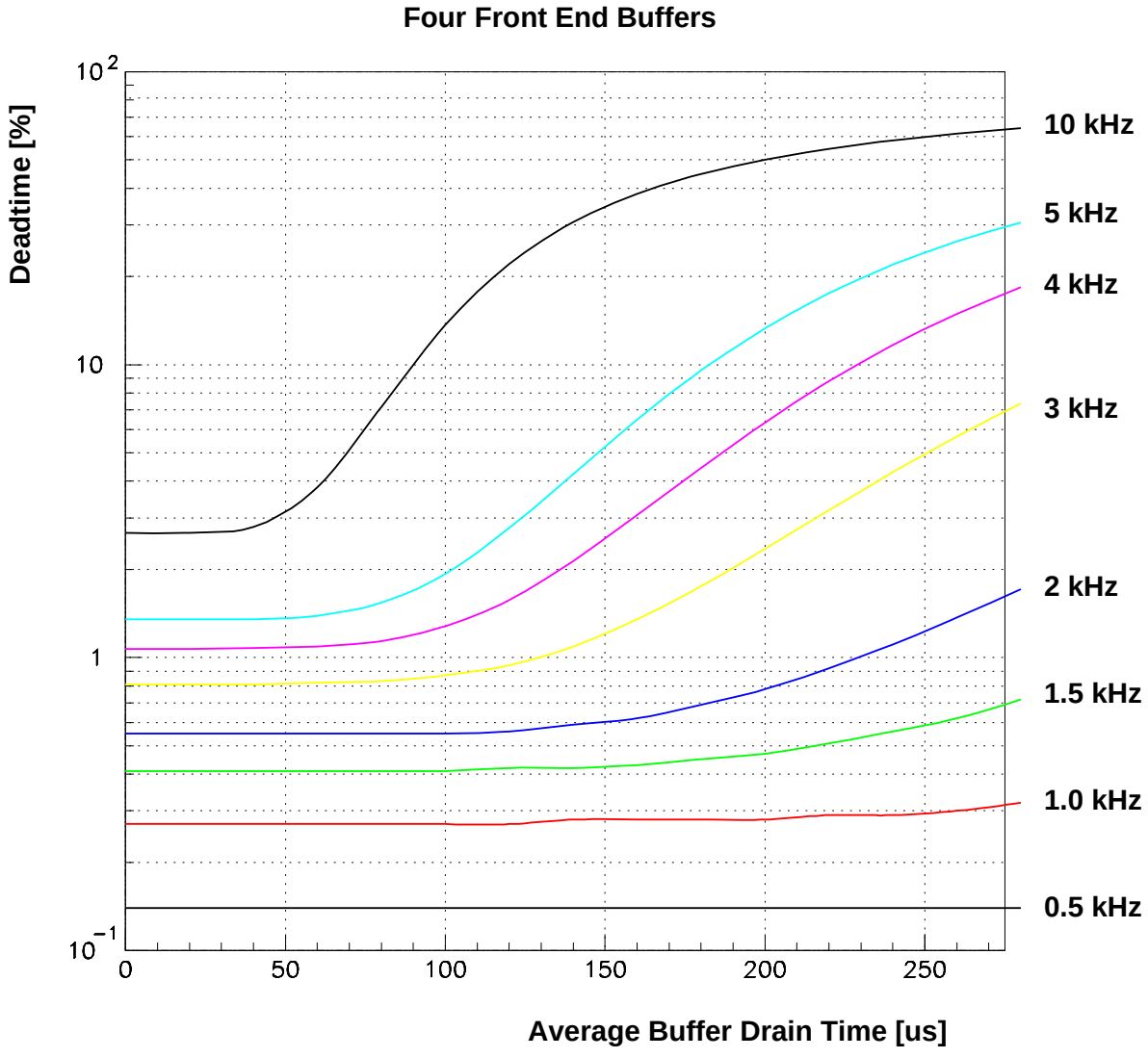
Bottlenecks

After the upgrades we have done so far, limits in the future will be:

- time required by the fragment level to send out data;
- time required to transport data from the fragment level to the UNIX nodes;
- amount of data sent on the drift chamber fiber link;
- feature extraction code for drift chamber and calorimeter.

FEE Buffer Model

- The model in the previous slides doesn't describe the FEE buffers.
- Our FEE can store up to 4 events, i.e. it's possible to give four L1 accepts before we need to read out events.
- Instrumental in reducing dead time when there are few events closely spaced in time.



- Horizontal axis is the fiber link column in model.
- Dead time simulated assuming a Poisson distribution.

Upgrade Plan (1)

- What we have done so far regarding the upgrades:
  - deployed VMON to understand system performance;
  - deployed software to allow system to run with an arbitrary number of L3 nodes (>32);
  - deployed software that allows ODF to run on Linux, which will make easier for the UNIX event builder and the trigger to run faster for less money;
  - deployed event batching which improves performance by grouping more events together to reduce overhead;
  - split some of the crates into smaller pieces to increase bandwidth and CPU.
- For the future upgrades we don't want to over design the system too much, but be somewhat conservative.
- We think the option of redesigning/rebuilding the existing system is too conservative. This would require a large engineering effort we think we cannot justify at this time given the headroom we have and the headroom we believe we can gain from upgrading the system.
- Our recommended approach until spring 2003 is to tweak the existing system to improve performance.

Upgrade Plan (2)

- This approach has significant advantages:
  - it's not too expensive (yet) if we can delay purchasing faster CPUs (which we believe to be the case);
  - it gives the opportunity to see what the outcome of the trigger upgrade is.
- We are considering these upgrades for the current system:
  - gigabit Ethernet with custom dataflow driver; current VxWorks driver has 166 μs overhead; the fact ODF uses UDP instead of TCP will significantly simplify the coding of the transport layer of the driver;
  - split drift chamber data over more fibers, or modify FEE thresholds to reduce data volumes;
  - improve FEX software for data coming from the calorimeter.
- On the longer term:
  - replace CPUs on all ROMs;
  - change bus technology between the segment and the fragment levels;
  - or getting rid of of the two stage event builder (i.e. removing the fragment level); this approach would have the downsides of removing a natural hierarchy, of generating more interrupts in the UNIX event builder and of leaning more on the switch.