Explorations of the viability of ARM and Xeon Phi for physics processing

Peter Elmer -
Princeton University

David Abdurachmanov -
Vilnius University

Giulio Eulisse,
Chris Jones,
Shahzad Muzaffar - *FNAL*

Kapil Arya,
Gene Cooperman -
Northeastern University

Tommaso Boccali -
INFN-Pisa

Andrea Dotti - *SLAC*

Francesco Giacomini,
Matteo Manzali - *CNAF*

Josh Bendavid - *CERN*

New architectures

- Over the past ten years processors have hit power limitations which place significant constraints on "Moore's Law" scaling.
- The first casualty was scaling for single sequential applications, giving birth to multi-core processors.

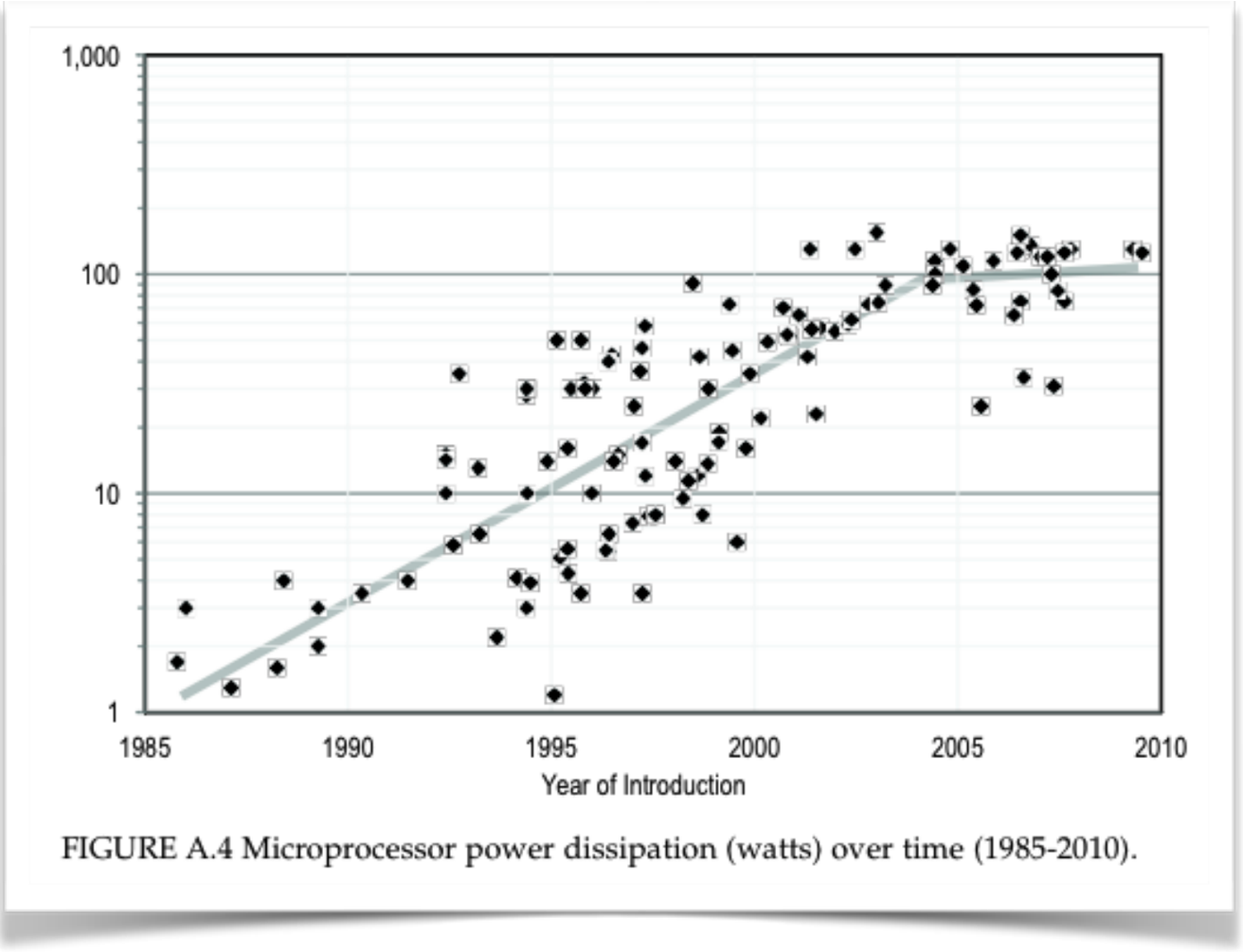
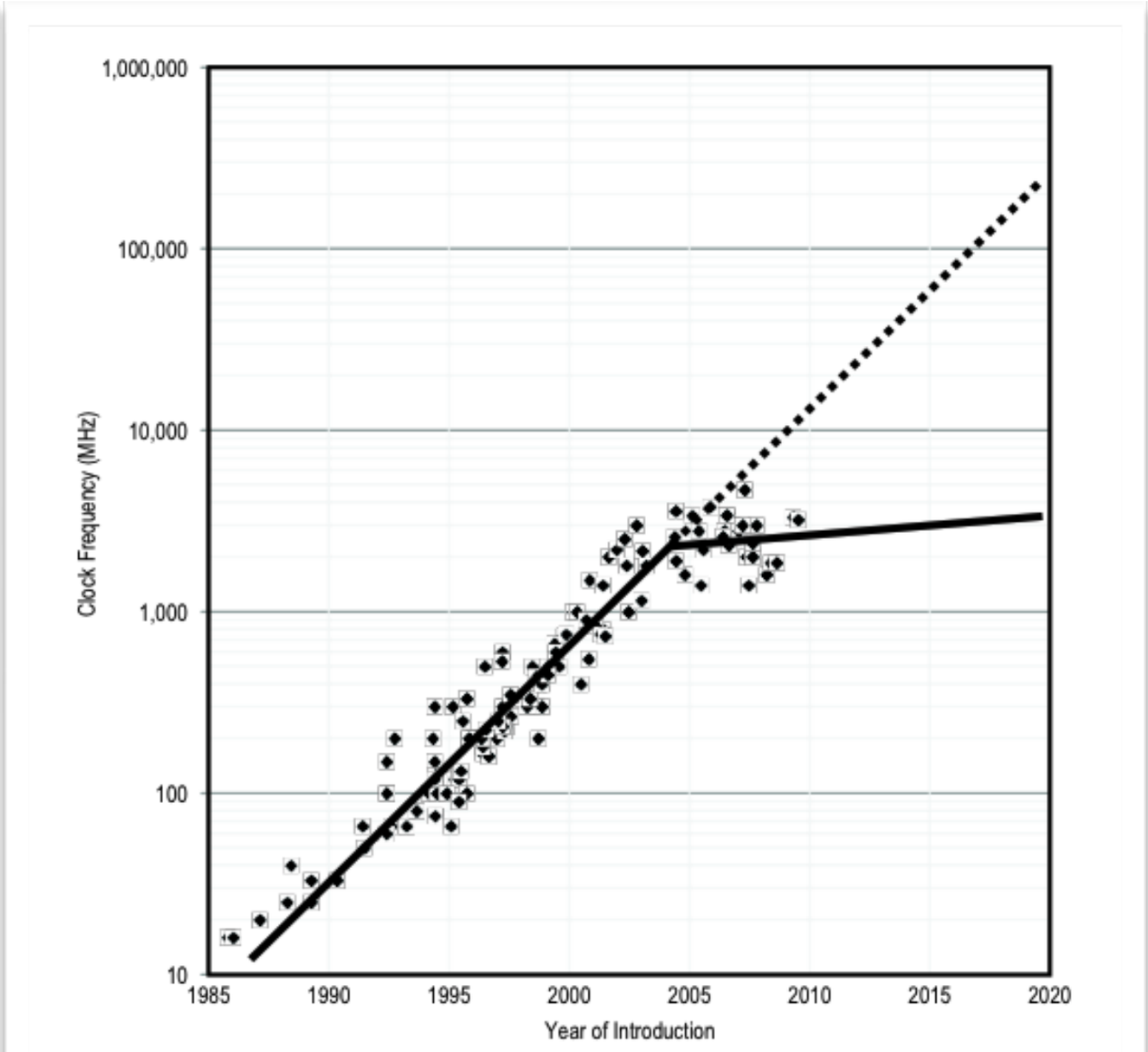


FIGURE A.4 Microprocessor power dissipation (watts) over time (1985-2010).

From: "The Future of Computing Performance: Game Over or Next Level?"

New Architectures

- Even multi-core, implemented with large "aggressive" cores is just a stop-gap. The power limitations remain. The focus is shifting to performance/watt, not just performance/price.



From: "The Future of Computing Performance: Game Over or Next Level?"

CMSDIST/PKGTOOLS

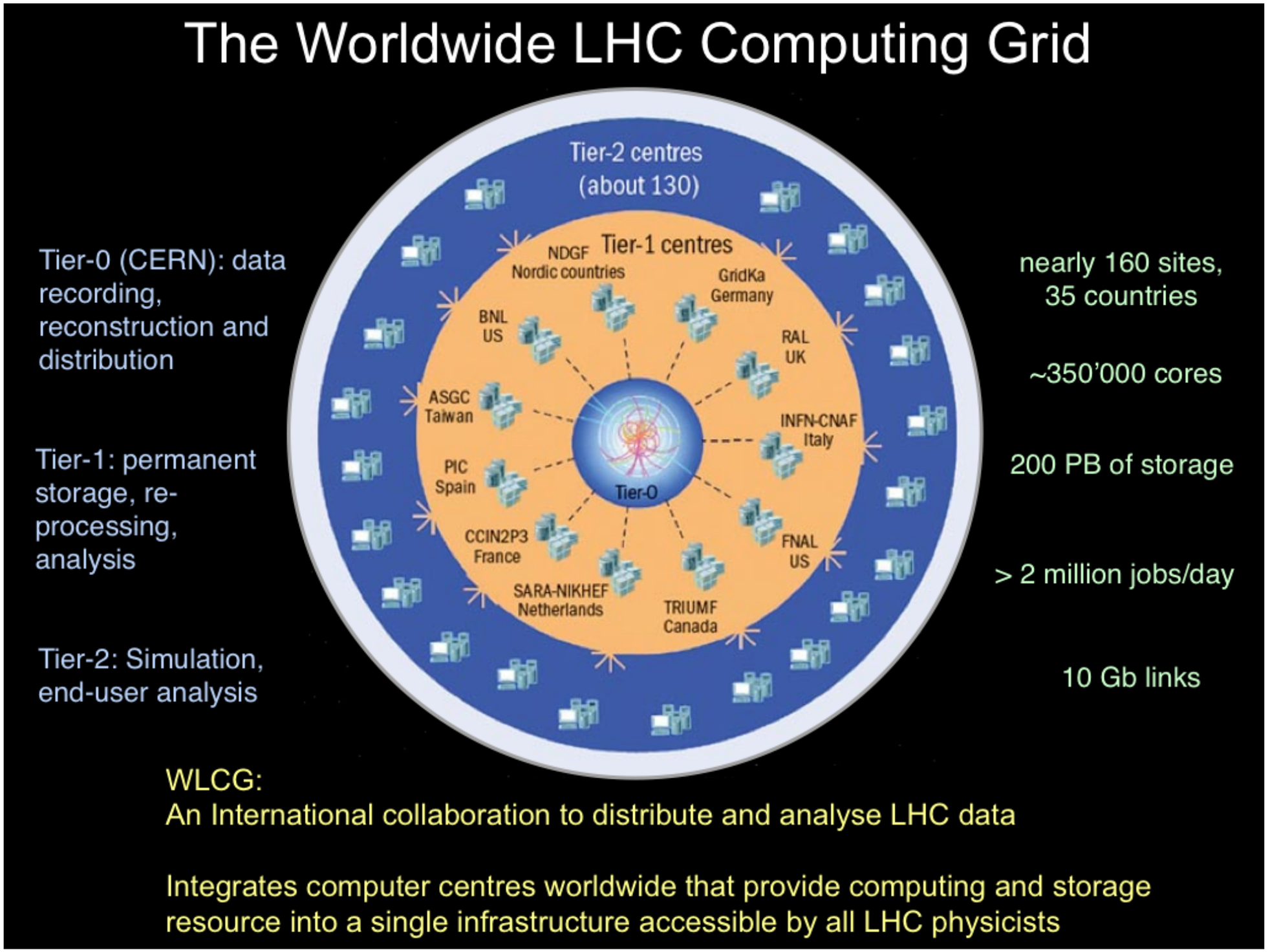
- In the following I will be talking about our ports of the CMS software and the stack of "externals" to ARMv7 and to the Xeon Phi, as well as some initial tests.

- Note that we don't just hack together some build locally on a machine for this. We use same set of build tools we use for x86-64. The build recipes are thus documented in the form of rpm spec files and a successful build results also in an rpm which can be uploaded to our apt repository.

- Build recipes: https://github.com/cms-sw/cmsdist

- Build tools: https://github.com/cms-sw/pkgtools

WLCG as Distributed Supercomputer



WLCG as Distributed Supercomputer - Power

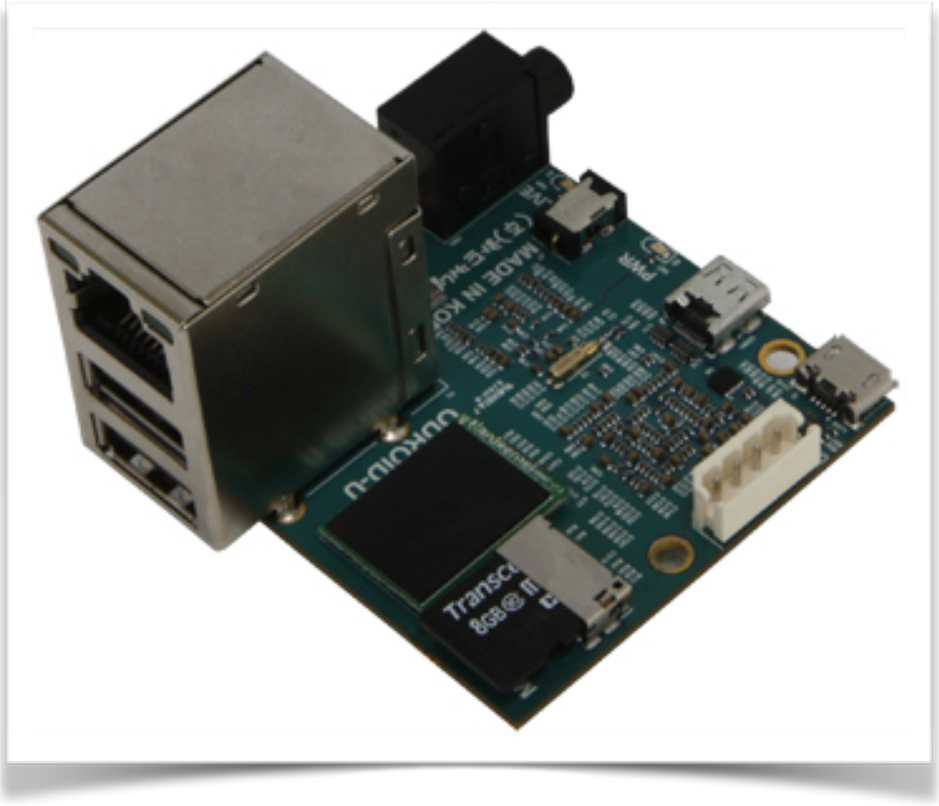
- Not only would the the WLCG be one of the top supercomputers in terms of performance if it were considered as such, but it also shares another characteristic which is less obvious.

- Using the mix of hardware available at FNAL (and known power use), we estimate the aggregate power cost to be of order 10MW

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National University of Defense Technology China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3120000	33862.7	54902.4	17808
2	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560640	17590.0	27112.5	8209
3	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1572864	17173.2	20132.7	7890
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705024	10510.0	11280.4	12660
5	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786432	8586.6	10066.3	3945
6	Texas Advanced Computing Center/Univ. of Texas United States	**Stampede** - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi SE10P Dell	462462	5168.1	8520.1	4510
7	Forschungszentrum Juelich (FZJ) Germany	**JUQUEEN** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	458752	5008.9	5872.0	2301
8	DOE/NNSA/LLNL United States	**Vulcan** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	393216	4293.3	5033.2	1972
9	Leibniz Rechenzentrum Germany	**SuperMUC** - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR IBM	147456	2897.0	3185.1	3423
10	National Supercomputing Center in Tianjin China	**Tianhe-1A** - NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 NUDT	186368	2566.0	4701.0	4040

ODROID-U2 Development Board

- Initial tests done with a small 32bit/ ARMv7 development board
- Exynos4412 Prime CPU
- 1.7GHz Cortex-A9 quad core
- 2GB L-DDR memory (total)
- eMMC, microSD, 2xUSB2.0, 10/100Mbps Ethernet
- $89 (~$233 with cables, cooling fan, 64GB eMMC, power adaptor, ...)

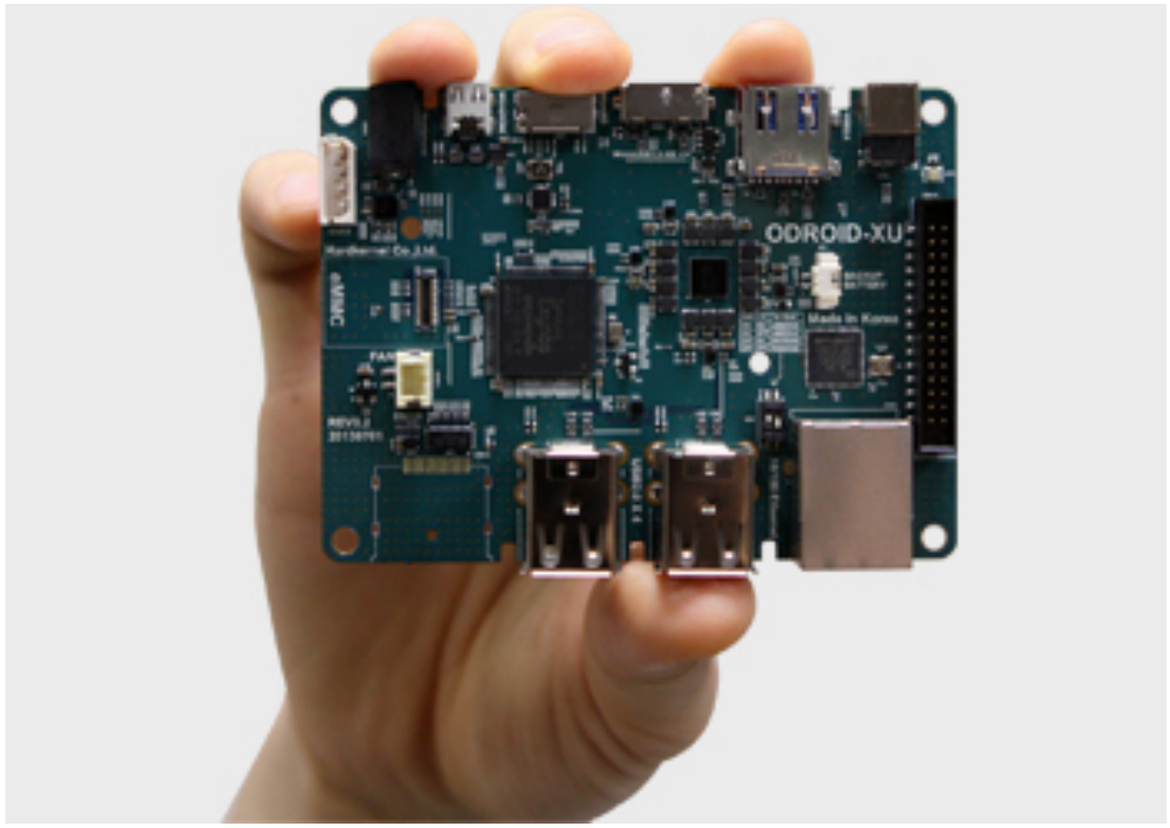


- Fedora 18 ARMv7-A, hard floats, gcc 4.8, ODROID kernel

Arrived Feb, 2013

ODROID-XU+E Development Board

- Exynos5 octa (5410) CPU
- 1.6GHz Cortex-A15 quad core + 1.2GHz Cortex-A7 quad-core (big.LITTLE heterogeneous mix)
- 2GB L-DDR memory (total)
- eMMC, microSD, 4xUSB2.0, 1xUSB3.0, 1xUSB3.0 OTG, 10/100Mbps Ethernet
- $199 (~$357 with 64GB eMMC, power adaptor, "Smart Power" meter ...)



- Fedora 19 ARMv7-A, hard floats, gcc 4.8, ODROID kernel

Arrived 25 Sep, 2013

ARM integration build

- For new architectures our aim has been not only to do an initial port of the software, but also to keep the build alive over time (and subsequent changes)
- For this reason we aim to add new ports to the daily integration builds (e.g. ARMv7 port)
- 16 packages (of 1106) don't build or link due to missing Oracle for ARM

Summary for CMSSW_7_0_X_2013-10-09-0200 IB on platform fc18_armv7hl_gcc481 -- Back to IB portal

error type	# of packages	total # of errors
dictError	0	0
compError	13	357
linkError	17	49
pythonError	0	0
compWarning	15	25
dwnlError	0	0
miscError	3	7
ignoreWarning	6	12
scram errors	0	unknown
scram warnings	0	unknown
libChecker	0	unknown

Log file from the BuildManager

- Log file from the BuildManager (check here if something _completly_ fails).
- Log file from "scram p" and CVS checkout.
- Log file from "scram b".

For the new libchecker errors and the SCRAM errors and warnings please click on the linked number to see the details for the package.

#/status	subsystem/package	compError	linkError	compWarning	miscError	UnitTest logfile
0	CalibCalorimetry/EcalPedestalOffsets	1	1	-	-	-
1	CalibCalorimetry/EcalSRTools	1	1	-	-	-
2	CalibCalorimetry/EcalTPGTools	4	1	-	-	-
3	CalibTracker/SiStripDCS	5	2	-	-	-
4	CaloOnlineTools/HcalOnlineDb	18	2	-	1	-
5	CondTools/Ecal	57	2	2	1	-
6	CondTools/Hcal	2	2	-	5	-
7	DQM/SiStripCommissioningDbClients	13	2	-	-	-
8	OnlineDB/CSCCondDB	29	6	-	-	-
9	OnlineDB/EcalCondDB	211	1	-	-	-
10	OnlineDB/SiStripConfigDb	9	3	-	-	-
11	OnlineDB/SiStripESSources	6	2	-	-	-
12	OnlineDB/SiStripO2O	1	1	-	-	-
13	CalibCalorimetry/EcalTrivialCondModules	-	1	-	-	-
14	Calibration/EcalCalibAlgos	-	2	-	-	-
15	Calibration/HcalCalibAlgos	-	1	-	-	-
16	CondCore/EcalPlugins	-	19	-	-	-
17	Calibration/HcalAlCaRecoProducers	-	-	2	-	-
18	Calibration/Tools	-	-	1	-	-
19	DataFormats/GeometrySurface	-	-	1	-	-
20	Fireworks/SimData	-	-	1	-	-
21	GeneratorInterface/HijingInterface	-	-	1	-	-
22	JetMETCorrections/IsolatedParticles	-	-	2	-	-
23	RecoLocalMuon/CSCEfficiency	-	-	1	-	-
24	RecoMuon/MuonIdentification	-	-	1	-	-
25	RecoMuon/MuonIsolation	-	-	2	-	-
26	SimGeneral/PileupInformation	-	-	1	-	-
27	TrackingTools/AnalyticalJacobians	-	-	3	-	-
28	TrackingTools/GsfTracking	-	-	1	-	-
29	TrackingTools/TrackAssociator	-	-	3	-	-
30	Utilities/Timing	-	-	3	-	-
31	Alignment/CocoaAnalysis	-	-	-	-	-
32	Alignment/CocoaApplication	-	-	-	-	-
33	Alignment/CocoaDDLObjects	-	-	-	-	-

See also "The Rise of the Build Infrastructure" (G.Eulisse) on Thursday

Status of ARMv7 port

- Everything basically builds except a small set of CMSSW packages which depend on Oracle

- At run time we still have some issue related to the ROOT dictionaries which prevents reading and writing CMS event data files. (This is after a couple of previous bugs were found and fixed, plus some amount of adding missing entries to dictionaries, as a follow-on to one of the fixes.)

- Some difficulties using (U2) Fedora18 builds on (XU+E) Fedora19, so we've been doing separate builds for now.

ODROID Power

- A comparison of the power cost of this small development board with a full server is a bit misleading, so we've tried to get numbers that correspond to the Thermal Design Power (TDP), for which there are published specs for x86-64.

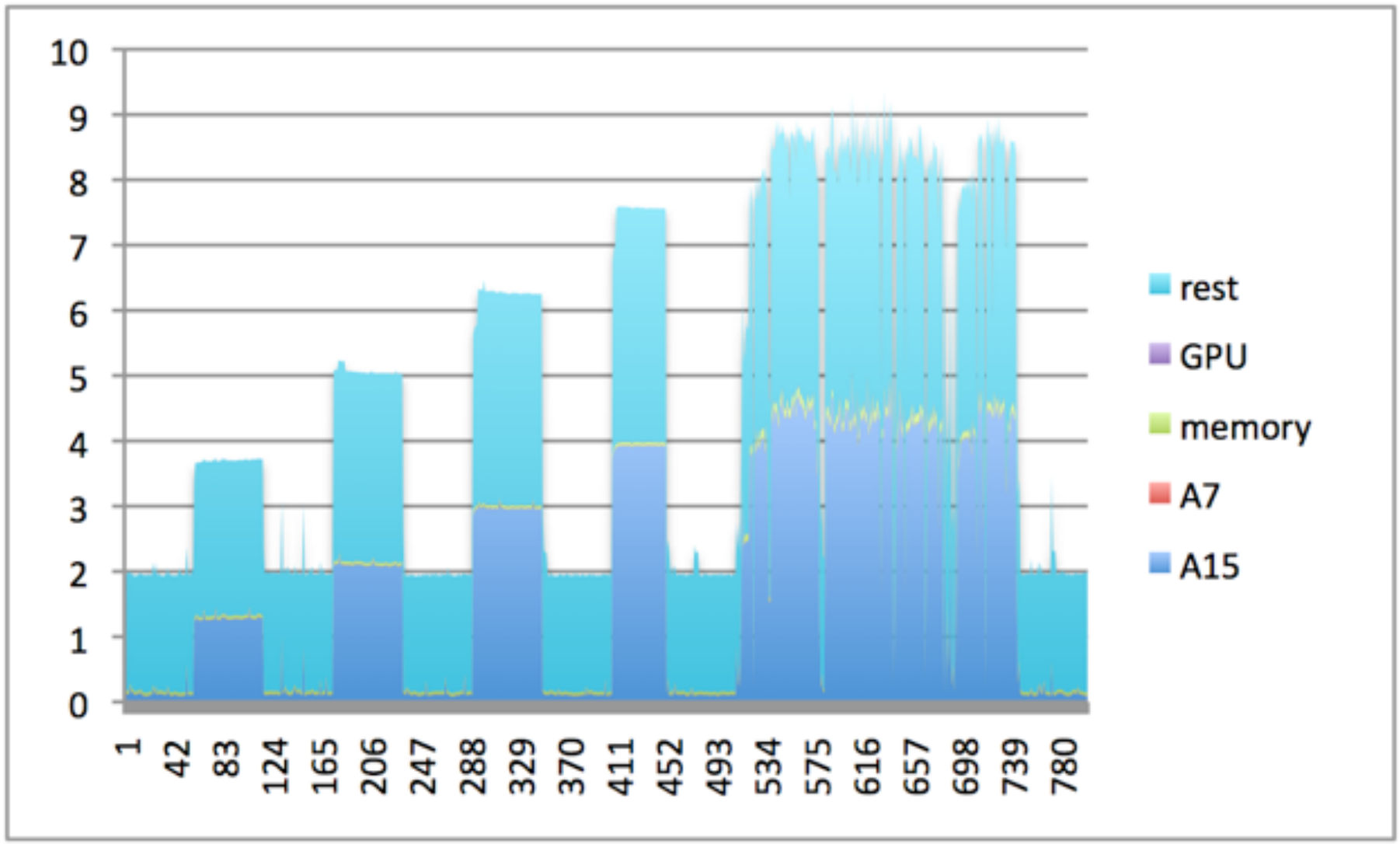
- For both the U2 and the XU+E boards we used an external "Smart Power" meter that also provides the measurements to the board via USB.

- For the XU+E board, there are also dedicated sensors that allow measurement of the power used by the A15 cluster, the A7 cluster, the memory and the GPU (independently)

ODROID U2 power (from "Smart Power" meter)

ODROID U2	Voltage (V)	Current (mA)	Power (W)
idle no fan (no eth)	5.02	280	1.4
idle w/ fan (no eth)	5.02	360	1.8
idle no fan (eth)	5.02	322	1.6
idle w/ fan (eth)	5.02	400	2.0
full CPU load no fan (eth)	5.02	900	4.5
full CPU load w/ fan (eth)	5.02	970	4.9

ODROID XU+E Power (Sensors and "Smart Power")



- Load (1,2,3,4) cores and the a compilation test while monitoring power (Watts versus time)

ARM Performance

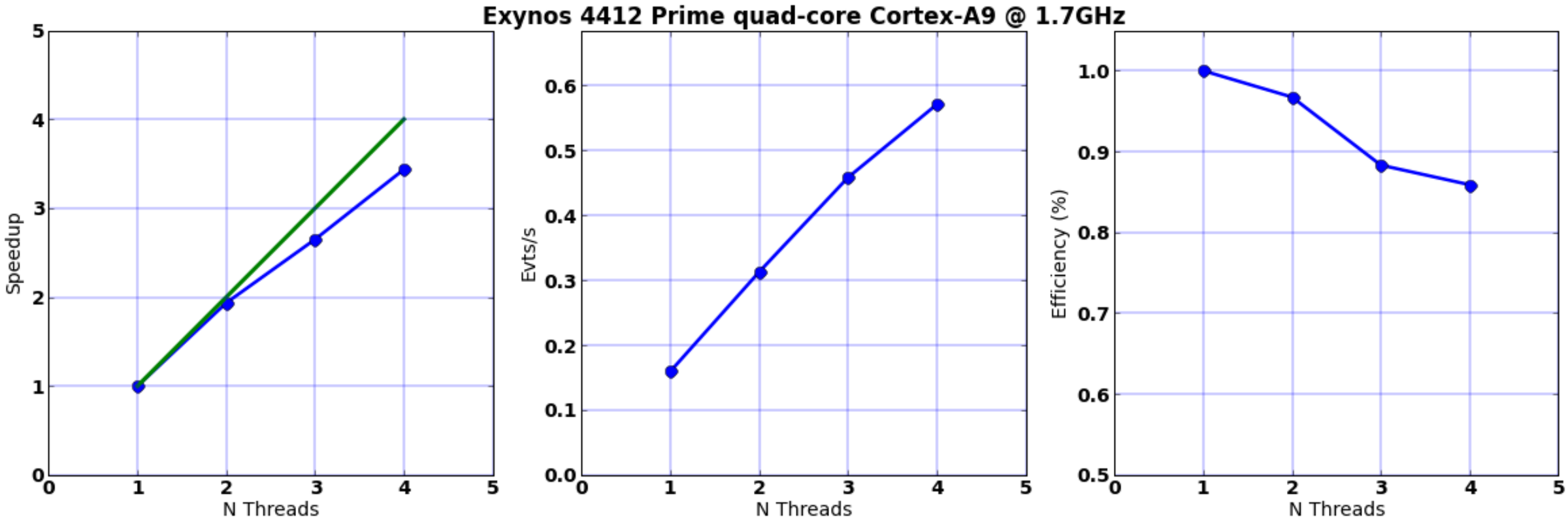
- We are running two types of tests:

- A full CMS job doing event generation and simulation ("GEN-SIM") of Minimum Bias events, with ROOT output turned off (due to remaining dictionary issues) - We run in single threaded mode and extrapolate to 4 cores, due to 2GB memory constraint

- Multithreaded Geant4 (Geant4-MT) "FullCMS" benchmark, with single pions - This fits easily in the memory and we can test vs #cores

	Cores	TDP	Gen-Sim Evt/min/ core	Gen-Sim Evt/min/ W	G4MT Evt/min (threads)	G4MT Evt/min/ W
ODROID U2	4	4W	1.14	1.14	34.2 (4)	8.6
ODROID XU+E	4/4	5.5W?			45 (4) (est.)	8.2
dual Xeon L5520	2x4	120W	3.50	0.23	307.2 (16)	2.6
dual Xeon E5-2630L	2x6	190W	3.33	0.21		

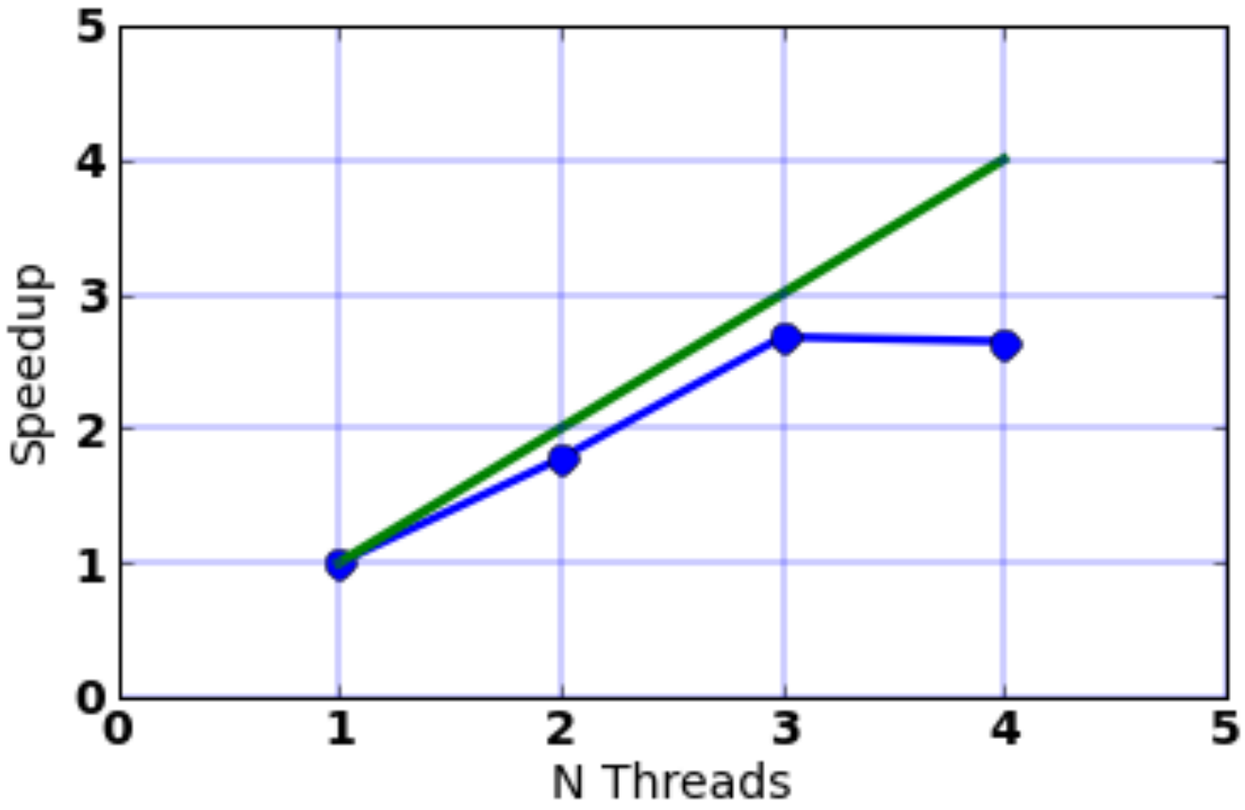
ARM Performance - Scalability vs #cores



- Use Geant4-MT (FullCMS) and run with 1 thread, 2 threads, 3 threads, 4 threads, on ODROID U2 board

- See also "Geant4 - Towards major release 10" talk (G.Cosmo) on Thursday. (For G4, these are preliminary numbers!)

Scalability test on ODROID XU+E



- Again use Geant4-MT (FullCMS) and run with 1 thread, 2 threads, 3 threads, 4 threads.

- Here one core turned itself off (overheating?), still under investigation. For full rate, we extrapolate from 3 to 4 cores.

IgProf on ARMv7/32bit - http://igprof.org

- We now have basic support for IgProf on ARMv7, memory profiling works, still debugging performance profiling

igprof-arm-MEM_TOTAL-10Evt-GEN-SingleElectron - navigator, performance

Back to profiles index

Counter: MEM_TOTAL, first 1000 entries

Sorted by self cost

Sorted by self cost

Sort by cumulative cost

Rank	Total %	Self	Calls	Symbol name
7	38.49	201,855,046	3,706,206	std::basic_string<char, std::char_traits<char>, std::allocator<char> >::_Rep::_S_create(unsigned int, unsigned int, std::allocator<char> const&)
10	33.45	175,397,348	5,026	frontierMemData_create
109	2.36	12,352,400	57,948	TString::Replace(int, int, char const*, int)
128	1.82	9,522,528	48,826	TString::Init(int, int)
37	1.36	7,135,467	114,275	G__search_tagname
159	1.35	7,078,336	28	PyObject_Malloc
221	1.00	5,228,000	2,517	XML_GetBuffer
224	0.99	5,186,832	5,026	poolGrow
233	0.94	4,935,532	10,052	parserCreate
240	0.91	4,753,588	87,207	TStorage::ObjectAlloc(unsigned int)
284	0.72	3,775,072	3,548	sre_match
295	0.66	3,457,344	5,287	dictresize
299	0.64	3,376,176	22,812	G__memfunc_next
303	0.63	3,280,896	801	G__strip_quotation
310	0.58	3,025,440	8,404	__fopen_internal
340	0.49	2,593,416	45,234	lookup

DMTCP

- Distributed MultiThreaded CheckPointing package (DMTCP), developed at Northeastern University (NEU), http://dmtcp.sourceforge.net

Key Features

Userspace checkpointing, no kernel-level access required

Checkpoints multithreaded applications

Checkpoints distributed applications

Can handle fork, exec, ssh, open file descriptors, TCP/IP sockets, etc.

Minimum runtime overhead

Optional compression of checkpoint images

Open source

Works on linux and supports a wide range of kernels

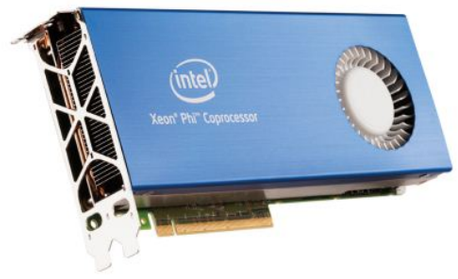
DMTCP 2.0

- DMTCP version 2.0 supports both application initiated checkpoints and, and via the use of plugins, flexible detach and reconnect to external resources.

- ARMv7 is also supported, along with x86-64, Intel MIC

Xeon Phi (7110P)

- 61 in-order lightweight cores with big vector units, coprocessor packaging on PCIe bus, 16GB GDDR5 memory



- Practical difficulties even to play with it:
  - Cross compilation from x86-64 required
  - Intel compiler required
  - No software environment available
- Offload vs direct running on the card (future?)

Xeon Phi

- Probably not sensible (or performant) to run entire CMS applications on the Xeon Phi, but it would facilitate tests to have some software development environment

- Solution: produce a CMSSW release subset with a smaller set of externals available (cross-compiled) and a subset of the CMSSW itself (code which compiles with icc)

- Mechanically it implies that one can create a SCRAM development area on the x86-64 host and code checked out and built will automatically be cross compiled for the Xeon Phi

Intel Compiler

- We began using icc 13.1.3, which had various problems with C++11. We then switched to icc 14.0.0, where the C++11 issues are resolved, but we were stopped by:

- http://software.intel.com/en-us/forums/topic/472385

- Aside: our experience with the Intel compiler has never been very good (going back ~10 years!) It lives up to its reputation as only a "benchmark compiler": compilation of real C++ codes almost always reveals non-compliance bugs, crashes, etc. Even if the bug above is fixed, we are at the mercy of Intel's release schedule (unlike gcc, where we can patch and move on).

Xeon Phi - CMSSW release subset status

- Release CMSSW_7_0_0_pre5 for slc6_mic_gcc481 architecture is available

- Externals: zlib bz2lib openssl expat readline sqlite db4 gdbm autotools python pcre xz libjpg xerces-c gccxml gsl ncurses pacparser photos pythia6 libpng libxml2 freetype lhapdf gmake cppunit fftw3 libuuid libtiff frontier_client xrootd boost clhep hepmc root fastjet heppdt nspr vdt sigcpp pythia8 tauola geant4 charybdis herwig classlib elementtree roofit coral (not all 125 yet, though)

- Should help facilitate some types of prototyping on Xeon Phi, without having to start from main() and by-hand Makefile

Cross-compilation notes

- Mostly external were build by just setting CXX="icpc -fPIC -mmic" and CC="icc -fPIC -mmic" and --host=x86_64-k1om-linux to configure scripts for cross compilation
- Boost: Patched couple of files [b] and used TOOLSET intel
- Python: Needs to build it twice, once for build system and once for Xeon Phi cross compilation.
- Fastjet: without -msse3
- GSL: Fixed configure script to not run test when cross-compiling
- OpenSSL: Configured w/o -fstack-protector and --with-krb5-flavor
- Root: Patched to build few executables without -mmic, Built without fftw3,castor and dcap dependency. Configured for linuxx8664k1omicc along with couple of patches to use freetype and pcre from cms externals. Pass -mmic option to icc fortran compiler

Xeon Phi Integration Build

- To avoid regressions relative to compilation with the Intel compiler, and facilitate testing with the latest code we have also set up an integration build for the Xeon Phi.

- 369 out of 1106 packages compile.

Summary for CMSSW_7_0_X_2013-10-10-0200 IB on platform slc6_mic_gcc481 -- Back to IB portal

error type	# of packages	total # of errors
dictError	0	0
compError	555	3718
linkError	734	1569
pythonError	0	0
compWarning	0	0
dwnlError	0	0
miscError	2	3
ignoreWarning	0	0
scram errors	0	unknown
scram warnings	0	unknown
libChecker	0	unknown

Log file from the BuildManager

- Log file from the BuildManager (check here if something completly fails).
- Log file from "scram p" and CVS checkout.
- Log file from "scram b".

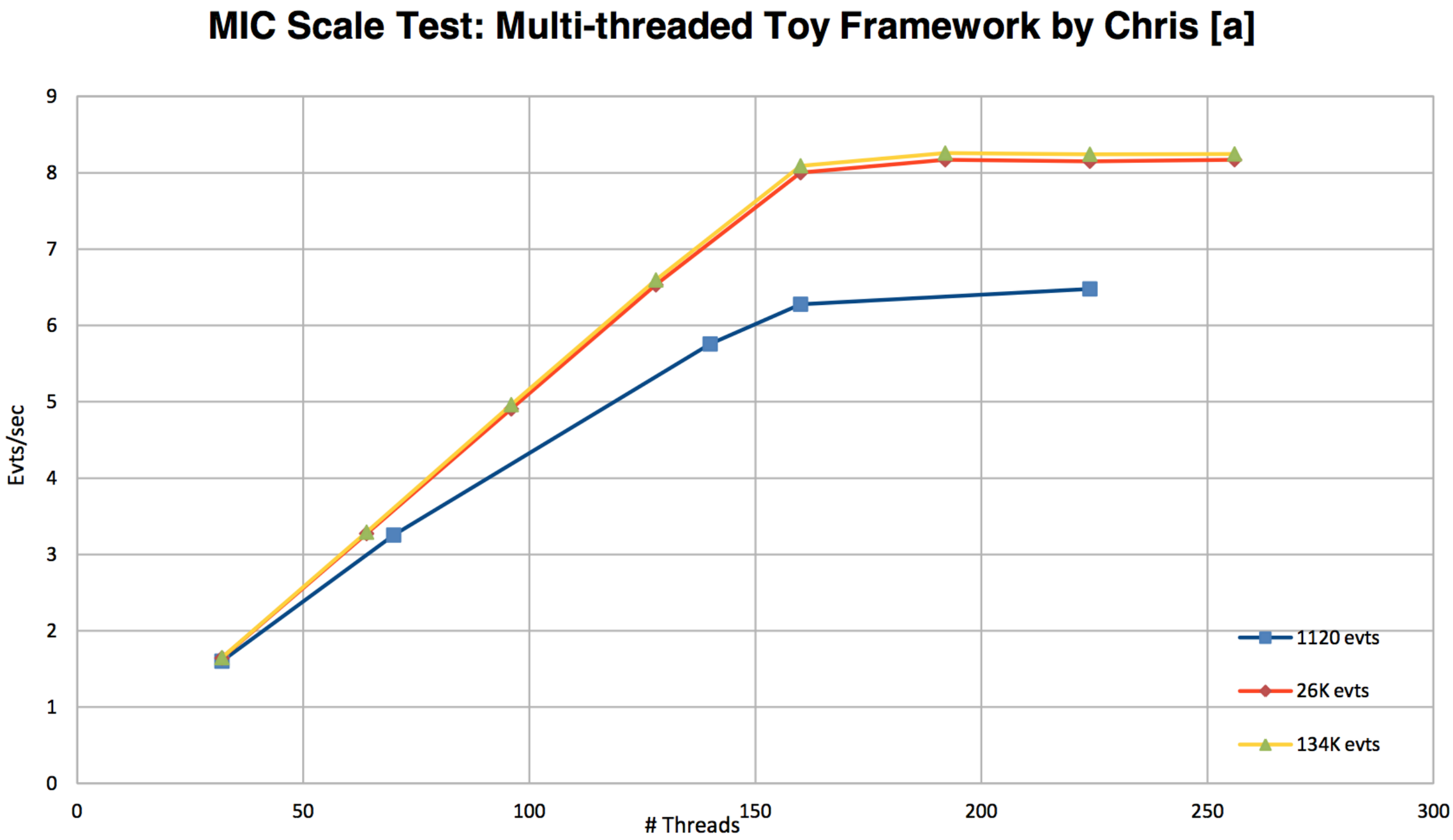
For the new libchecker errors and the SCRAM errors and warnings please click on the linked number to see the details for the package.

#/status	subsystem/package	compError	linkError	miscError	UnitTest logfile
0	Alignment/CocoaApplication	1	1	-	-
1	Alignment/CocoaFit	1	1	-	-
2	Alignment/CommonAlignment	11	1	-	-
3	Alignment/CommonAlignmentMonitor	9	11	-	-
4	Alignment/CommonAlignmentParametrization	7	1	-	-
5	Alignment/Geners	15	1	-	-
6	Alignment/HIPAlignmentAlgorithm	3	1	-	-
7	Alignment/LaserAlignmentSimulation	8	5	-	-
8	Alignment/LaserDQM	2	1	-	-
9	Alignment/MuonAlignment	29	7	-	-
10	Alignment/MuonAlignmentAlgorithms	28	6	-	-
11	Alignment/SurveyAnalysis	18	4	-	-
12	Alignment/TwoBodyDecay	1	1	-	-
13	AnalysisAlgos/TrackInfoProducer	4	3	-	-
14	AnalysisDataFormats/EWK	3	2	-	-

737	Utilities/LStoreAdaptor	-	2	-	-
738	HLTrigger/Tools	-	-	1	-
739	RecoCaloTools/MetaCollections	-	-	2	-
740	Alignment/CocoaDDLObjects	-	-	-	-
741	Alignment/CocoaDaq	-	-	-	-
742	Alignment/CocoaUtilities	-	-	-	-
743	Alignment/CommonAlignmentAlgorithm	-	-	-	-
744	Alignment/CommonAlignmentProducer	-	-	-	-
745	Alignment/KalmanAlignmentAlgorithm	-	-	-	-

1098	Validation/RecoB	-	-	-	-
1099	Validation/RecoEgamma	-	-	-	-
1100	Validation/RecoMuon	-	-	-	-
1101	Validation/RecoPixelVertexing	-	-	-	-
1102	Validation/RecoTrack	-	-	-	-
1103	Validation/RecoVertex	-	-	-	-
1104	Validation/Tools	-	-	-	-
1105	Validation/TrackerConfiguration	-	-	-	-
1106	Validation/TrackerRecHits	-	-	-	-

Toy Framework Scale Test



Real CMS code with RooFit and OpenMP

- As a test with real code, we took a fitting example: photon energy regression training (for Higgs->gamma gamma) via a ROOT macro

- The macro uses a compiled library (a Gradient Boosted Regression (GBR) Likelihood implementation) and some steering code in the form of a ROOT macro (using Roofit). On a more standard x86 machine, the macro is run using ACLiC ROOT to compile the steering code, and then call the library.

- This turned out as not possible on the MIC, since ACLiC needs a compiler, which we had only installed on the hosting machine; we had hence to revert to fully compiled code. OpenMP is standard mode starts a number of threads equal to the number of logical cores the system sees, 243 in our case.

- While resulsts are not conclusive at the moment, we have been able to see machine load level as high as 20% (via micsmc).

Summary

- We have put together software ports for both small ARMv7 development boards and a basic software environment for testing the Xeon Phi

- We have obtained some first benchmarks of applications on the ARMv7 processors and have begun to run and test applications on the Xeon Phi.