Virtual Institute – High Productivity Supercomputing

VI-HPS

MAC / VI-HPS Workshop Profiling & Performance Analysis of Parallel Applications KAUST, Saudi Arabia

October/November 2010

an Research Sch

		_
11	\square	
T	ТТ	
Ш		
Ш		

ÜLICH

Munich Centre for Advanced Computing

جامعة الملك عبدالله للعلوم والتقنية King Abdullah University of Science and Technology



TECHNISCHE

TENNESSEE

Universität Stuttgart



- Presenters/Guides
 - Hans-Joachim Bungartz, Michael Gerndt, Josef Weidendorfer, Tobias Weinzierl (TU Munich)
 - Matthias Weber (TU Dresden ZIH)
 - Felix Wolf (GRS-Sim)
 - Brian Wylie (JSC)
- Thanks
 - VI-HPS/POINT partners & associates
 - U Oregon PRL for preparing the LiveDVD
 - Local arrangements & facilities
 - ► KSL, KAUST

VI-HPS

We'd like to know a little about you, your application(s), and your expectations and desires from this tutorial

- What programming paradigms do you use in your app(s)?
 - only MPI, only OpenMP, mixed-mode/hybrid OpenMP/MPI, ...
 - Fortran, C, C++, mixed-language, ...
- What platforms/systems *must* your app(s) run well on?
 - Cray XT, IBM BlueGene/P, SGI Altix, Linux cluster[™], ...
- Who's already familiar with *serial* performance analysis?
 - Which tools have you used?
 - ► time, print/printf, prof/gprof, ...
- Who's already familiar with *parallel* performance analysis?
 - Which tools have you used?
 - ► time, print/printf, prof/gprof, mpiP/ompP, IBM HPC Toolkit, ...



- **Goal**: Improve the quality and accelerate the development process of complex simulation codes running on highly-parallel computer systems
- Funded by Helmholtz Association of German Research Centres



- Activities
 - Development and integration of HPC programming tools
 - Correctness checking & performance analysis
 - Training workshops
 - Service
 - Support email lists
 - Application engagement
 - Academic workshops

www.vi-hps.org

VI-HPS partners & associates





Forschungszentrum Jülich

RWTH Aachen University

- Jülich Supercomputing Centre



- Centre for Computing & Communication
 Technical University of Dresden
 - Centre for Information Services & HPC
- University of Tennessee (Knoxville)
 - Innovative Computing Laboratory



German Research School

- Laboratory of Parallel Programming
- **Technical University of Munich**
 - Chair for Computer Architecture



HPC Centre











Universität Stuttgart





- Marmot
 - Free MPI correctness checking tool
- PAPI
 - Free library interfacing to hardware performance counters
- Periscope
 - Prototype automatic analysis tool using an on-line distributed search for performance inefficiencies
- Scalasca
 - Open-source toolset for analysing the performance behaviour of parallel applications to automatically identify inefficiencies
- Vampir/VampirTrace
 - Commercial tool for graphical trace visualization & analysis, and open-source event tracing library

[Productivity Tools Live-DVD contains latest tools releases]

Marmot

VI-HPS

Tool to check for correct MPI usage at runtime

- Checks conformance to MPI standard
 - Supports Fortran & C bindings of MPI-1.2
- Checks parameters passed to MPI
- Monitors MPI resource usage

Implementation

- C++ library gets linked to the application
- Does not require source code modifications
- Additional process used as DebugServer
- Results written in a log file (ASCII/HTML/CUBE)

Developed by HLRS & TU Dresden

- Released as open-source
- http://www.hlrs.de/organization/av/amt/projects/marmot



Marmot logfiles

VI-HPS

	livetau(@localhost:Exer	rcise		_ + ×	Cube 3.2 QT	: Exercise/Marmot_data	type.exe_20090807_132838.cube	_ + X
<u>File Edit View Terminal Tabs Help</u>						<u>File D</u> isplay Jopology <u>H</u> elp			
1 (localhost.localdoma	in) mation see:/usr/local/packages/marmot-2.3				â	Absolute	✓ Absolute	~	Absolute
for MPI-Standard infor				ges/marmot-2.	3.0/share/doc/m	Metric tree	Call tree Fla	at view	System tree
armot-2.3.0/MPI-STANDARD/marmot err/node164.html						🖻 🗌 0 Messages	二 白 🗌 0	CallTree	🕒 🗔 - MPI-Enviroment 🔗
						- 📕 38 Infos		0 Unknown Location	🗗 🗌 - MPI-Processes
3: Warning global messa	ae with Tex	Text: Processes 0 and 1 both ru			n on localhost.	localhost 🗄 🖶 6 Warnings		0 Notes	- 🗌 0 rank 0
localdomain	ge witten tex		000000					0 Warning Toxtmoscages	
for MPT Standard infor	mation cool/		al /nacka	acc/marmot 2	3 Q/sharo/dos/m	L 0 Enors	valid	0 Frror Textmessages	
			ас/раска	iges/marmot-z.	5.0/ share/ uoc/ iii			0 datatype.c	
armot-2.3.0/MP1-STANDAR	D/marmot_err,	/node16	5.ntml					0 MPI Init @line: 47	
							-	0 MPI_Comm_rank @line	
10: Error from rank O(T	hread: 0) wi [.]	th Tex	t: ERROF	R: MPI_Send: d	atatype is not			0 MPI_Comm_size @line:	
valid!								0 MPI_Type_contiguous (
								■ I MPI_Recv @line: 56	
On Call: MPI Send From:	datatvpe.c	line: 5	3 for MF	PI-Standard in	formation see:/			0 MPI_Selid @illie: 55	h l
usr/local/packages/mar	0				MARMOT HTML Logfile - K	onqueror		= + X	
/node28_html	<u>File E</u> dit <u>V</u> iew <u>G</u> o <u>B</u> o	ookmarks <u>S</u> et	ttings <u>W</u> indow	<u>H</u> elp				2 <u>8</u>	
	4 × 4 × A × 🕫 6		nome/livetau/wo	rkshop-marmot/Exercise/Mai	mot datatype exe 20090807 130509 btm			a v -1	
10. Ennon from month 1/1	👾 🖤 🐨 🖶 😧 😭 😰 🕼 momenivetau/workshop-marmot/Exercise/Marmot_datatype exe_20090807_130509.html							v • .	
10: Error Trom rank 1(default: 1000 micros	seconds)		Â	
valid!					Text: MARMOT MAY				
			-				I		
On Call: MPI Recv From	0	Global	0	Information	(maximum message	time,	Unknown		
usr/local/packages/mar					default: 0 microseco	nds)			
/node28.html					Toxt MADMOT MAY	TMEOUT TWO = 0			
,					PIEXC MARMOT_MAX	$_1 \text{IMEOUT}_1 \text{WO} = 0$			
[livetau@localbost Evo	0	Global	0	Information	(maximum message	time,	Unknown		
					default [.] 0 microseco	nds)			-
					Taxts MADMOT LOC				
					lext: MARMOT_LOG	$FILE_PATH =$	Unknown		1 0 1 (100.00%) 1
	0	Global	0	Information	(path of Marmot log	file output,			
					default:)	•			
	0	Global	0	Information	lext: MARMOT_ERRO	$LODES_SET = (not set)$	Unknown		
	0	Giobai	U	mation	(not functional yet)		OTIKITOWIT		
	0	Global	0	Information	Text: End of the env	ironmental variables info	Unknown		
	0	Giubai	0	mormation	lext. End of the env		UTIKITUWIT		
	0	Clobal	0	Information Warning	lext: Thread Syncror	nisation is disabled.If you	Unknown		
	0	Giobal	0		are using multiple th	reads errors might occur	UNKIIOWII		
					Toxt: Dobuggoryer r	inc on camo nodo ac		Infoc coo	
	3	Global	0		lext. Debugserver n	ins on same noue as	Unknown	intos see	
		2.000			process 0 (localhost.	localdomain)	2	MPI-Standard	
	3 (_	Warning	Text: Debugserver ri	ins on same node as		Infos see	
		Global	0		process 1 (localbest	localdomain)	Unknown	MDI Standard	
					process I (localnost.	iocaldomain)		MPI-Standard	
	2		0	Manufactor	Text: Processes 0 and 1 both run on			Infos see	
	3	Global	0	warning	localbost localdomai	'n	Unknown	MPI-Standard	
								ministanduru	
					lext: ERROR: MPI_Se	end: datatype is not valid!	datatype c	Infoc coo	
	10	0	0	Error			udiacype.c	MDI CI	
	Call: MDL Sond			Call: MPL Sond		line: 53	MPI-Standard		
					Cuil. MFT_Sellu				
					Text: ERROR: MPI Re	ecv: datatype is not valid!	datatype c Infec coo		
	10 1 0 Error		_		uatatype.c	intos see			
	10	-	Ŭ		Call: MPL Book		line: 56	MPI-Standard	
								0	

VI-HPS

Portable performance counter library

- Configures and accesses hardware/system counters
- Predefined events derived from available native counters
- Core component for CPU/processor counters
 - instructions, floating point operations, branches predicted/taken, cache accesses/misses, TLB misses, cycles, stall cycles, ...
 - performs transparent multiplexing when required
- Extensible components for off-processor counters
 - ► InfiniBand network, Lustre filesystem, system hardware health, ...
- Used by multi-platform performance measurement tools
 - ► PerfSuite, Periscope, Scalasca, TAU, VampirTrace, ...

Developed by UTK-ICL

Available as open-source for most modern processors http://icl.cs.utk.edu/papi/



Periscope



Automated profile-based performance analysis

- Iterative on-line performance analysis
- Automatic search for bottlenecks based on properties formalizing expert knowledge
 - MPI wait states
 - Processor utilization hardware counters
- Multiple distributed hierarchical agents
- Eclipse-based integrated environment

Supports

SGI Altix Itanium2, IBM Power and x86-based architectures

Developed by TU Munich

- Released as open-source
- http://www.lrr.in.tum.de/periscope



Periscope plug-in to Eclipse



	6	Java - GENE/field_solve_l	kxky-psc.f90	- Eclipse Platf	orm		_ _ ×				
	Ele <u>E</u> dit <u>N</u> avigate Se <u>a</u> rch <u>P</u> roject <u>B</u> un <u>W</u> indow <u>H</u> elp										
	😫 Packag 😫 🁔 Hierarc 🖓 🗖	g_sca_128_install.ps 📄 comm-psc.f90 📄	field_solve_look	y-psc ដ "۱	_	0 20	Dutline 🔿 SIR Outline 😫 📑 Task List 💾 🗖				
	 ☐ ♣ ₽ ▼ ☐ fourier_mkl.F90 ♀ g_16_sca.psc ♀ g_16_sca.pscold ♀ g_16_scabf.psc ♀ g_16_scabf.pscold ♀ g_32_sca.psc 	33 Real, Dimension(:,:,:,:,:,:,:), Allo 34 Complex, Dimension(:,:,:), allocatab 35 36contains 37 Subroutine field_solve_kxky(p_g_1,p_4 38! Arguments 39 Complex, Dimension(li1:li2,lj1:lj2, 40 complex, dimension(lbx:ubx,lj1:lj2, 41 42! Local variables (put on stack) 43	catable:: mm lg :: p_phi_ emfields) ,lk1:lk2,ll1 ,lbz:ubz,l:n	4 4 4 4	 SIR File: /home/rainy/workspaces/runtime-Psc. subroutine: CALC_REST (54/220) (1-34) call: FIELD_SOLVE_KXKY (166/166) (1-42 subroutine: CALFULLHHS_KXKY_1 (40/119) call: MY_REAL_MAX_TO_ALL (79/79) (4-2 subroutine: MY_REAL_MAX_TO_ALL (58/172 call: MPLALLREDUCE (114/114) (10-240 subroutine: MY_COMPLEX_SUM_VWSPEC (4 						
So	g_32_sca.pscold g_32_scabf.psc g_32_scabf.psc g_32_scabf.pscold @n_64_sca_inst	<pre>4 Integer :: j, k, l, m, n, o complex, dimension(li1:li2,lj1:lj2, 47 Call perfor ('FldSolvesf') Sal Gyrpawerage and calculation of the fi </pre>	,lk1:lk2,l:n ,lk1:lk2,l:n	n_fields] :: mx n_fields,ln1:lr	uments 12) :: <u>vmoments</u> stribution fur	¢	Call: MPI_ALLRED UCE (406/406) (10-312 subroutine: FIELD_SOLVE_KXKY (131/320) (call: MY_CONVEX_SUM_VWSPEC (189/1 program: GENE (0 4) (18-21) ♥ loop: (0/334) (18 ♥ loop: (0/334) (18 ♥ loop: (203 loop: (203 loop: (203 loop: (203 loop: (213				
	<pre> g_B_sca_inst15.psc g_B_sca_inst18.psc g_B_sca_inst6.psc g_B_sca_inst8.psc g_B_sca_inst8.psc g_B_sca_psc1 g_B_sca.psc1 g_B_sca.psc1 g_B_scabf.psc g_B_scabf.psc g_B_scabf.psc1 g_B_scabf.psc2 g_B_scabf.psc3 g_B_scabf.psc3 g_B_scabf.psc3 g_B_scabf.psc3 g_B_scabf.psc4 g_B_scabf.psc4 g_B_scabf.psc4 g_B_scabf.psc5 g_B_sc</pre>	<pre>Sof Gyroaverage and catchallon of the 1. Sof Gyroaverage and catchallon of the 1. Sof Uthere is no routine for real*complex Sof Sof if (perf_vec(1).eg.1) then Call calc_moments(n_fields,.fals Sof else Call calc_moments_perf(lijk0,lln Call calc_moments_S) Call moments=sum(ymoments,S) Call my_complex_sum_vwspec(moments) Call my_complex_sum_vwspec(mo</pre>	g View Search Console SIR outlin SIR outlin SIR outlin SIR outlin SIR outlin SIR outlin SIR outlin	e view							
	GENE script sh	Name	Process S	Severity	Filename	Confidenc	ce Extra				
	gene-psc.f90	Stalls due to waiting for data delivery to register Stalls due to waiting for data delivery to register Stalls due to waiting for data delivery to register	46 5	30.22 field_s 30.32 field_s	olve_lodcy-psc.f90 olve_lodcy-psc.f90	1.00					
Pro	ject view	L2 misses	102	30.53 field_s	olve_kxky-psc.f90	1.00	es=221330 L2Misses=164831 L3Misses=				
	📄 gene128front 📄 gene512frontold 📄 gene512frontold1	IA64 Pipeline Stall Cycles	17 4 56	31.11 held_s 31.14 field_s 31.38 field_s	oive_koky-psc.f90 olve_koky-psc.f90 olve_koky-psc.f90	1.00 1.00 1.00		2			
	GeneFiles.txt Geneout.tar Current Sectors	IA64 Pipeline Stall Cycles IA64 Pipeline Stall Cycles Filter:		31.65 field_s 31.68 field_s 131.58 field_s 131.5 5ort: [Severity (FV	olve_loxky-psc.f90 olve_loxlov-psc.f90 hown - 1 Selected - /D)]	1.00	Properties view				

Scalasca

VI-HPS

Automatic performance analysis toolset

- Scalable performance analysis of large-scale applications
 - particularly focused on MPI & OpenMP paradigms
 - analysis of communication & synchronization overheads
- Automatic and manual instrumentation capabilities
- Runtime summarization and/or event trace analyses
- Automatic search of event traces for patterns of inefficiency
 - Scalable trace analysis based on parallel replay
- Interactive exploration GUI and algebra utilities for XML callpath profile analysis reports

Developed by JSC & GRS

- Released as open-source
- http://www.scalasca.org/



Scalasca automatic trace analysis report







Interactive event trace analysis

- Alternative & supplement to automatic trace analysis
- Visual presentation of dynamic runtime behaviour
 - event timeline chart for states & interactions of processes/threads
 - communication statistics, summaries & more
- Interactive browsing, zooming, selecting
 - Inked displays & statistics adapt to selected time interval (zoom)
 - scalable server runs in parallel to handle larger traces

Developed by TU Dresden ZIH

- Open-source VampirTrace library bundled with OpenMPI 1.3
- http://www.tu-dresden.de/zih/vampirtrace/
- Vampir Server & GUI offered with a commercial license
- http://www.vampir.eu/



Vampir interactive trace analysis GUI





Technologies and their integration







Key tool components also provided as open-source

- Program/library instrumentation
 - ► OPARI, POMP, PDToolkit
- MPI library/tool integration
 - ► UniMCI
- Scalable I/O
 - ► SIONIib
- Libraries & tools for handling (and converting) traces
 - ► EPILOG, EARL, PEARL, OTF
- Analysis algebra & hierarchical/topological presentation
 - ► CUBE

POINT/VI-HPS collaboration





VI-HPS collaborates with the POINT project in the USA

- Petascale Productivity from Open, Integrated Tools
- Funded by US NSF SDCI, Software Improvement & Support
- University of Oregon, University of Tennessee, UIUC NCSA, and Pittsburgh Supercomputing Center
- www.nic.uoregon.edu/point

VI-HPS

Entry-level (routine) profiling tools

- Intended to be simple to use, low measurement overhead
 - works with unmodified, dynamically-linked executables
- Statistical sampling profiles based on time or HWC events
- XML-based reports with configurable processing utilities
 - ► can be viewed in Web browser, with ParaProf or Cube3
- Processor inventory utility captures measurement metadata
- Performance event measurement configuration utility

Developed by UIUC/NCSA

- Available as open-source for x86, x86-64 & ia64 Linux
- Can be used with MPI, OpenMP & pthreads
- http://perfsuite.ncsa.uiuc.edu/



PerfSuite psconfig & performance reports







Integrated performance toolkit

- Instrumentation, measurement, analysis & visualization
 - Highly customizable installation, API, envvars & GUI
 - Supports multiple profiling & tracing capabilities
- Performance data management & data mining
- Targets all parallel programming/execution paradigms
 - Ported to a wide range of computer systems
- Performance problem solving framework for HPC
- Extensive bridges to/from other performance tools
 - ► PerfSuite, Scalasca, Vampir, ...

Developed by U. Oregon/PRL

- Broadly deployed open-source software
- http://tau.uoregon.edu/



TAU Performance System components

VI-HPS





- Goals
 - Give an overview of the programming tools suite
 - Explain the functionality of individual tools
 - Teach how to use the tools effectively
 - Offer hands-on experience and expert assistance using tools
 - Receive feedback from users to guide future development
- For best results, bring & analyse/tune your own code(s)!
- VI-HPS Tuning Workshop series
 - Aachen (3/08), Dresden (10/08), Jülich (2/09), Bremen (9/09), Garching (3/10), Amsterdam (05/10)
- Joint POINT/VI-HPS Tutorial series
 - SC (11/08), ICCS (5/09), SC (11/09), SC (11/10)
- Training with individual tools & platforms (e.g., BlueGene)



Sunday 31st October

- 08:30 (registration & notebook computer set-up)
- 09:00 Welcome [Bungartz, TUM]
- 09:15 Introduction [Wylie, JSC]
 - Virtual Institute High Productivity Supercomputing
 - Building and running the tutorial exercise NPB3.3-MPI/BT
- 09:45 Intro. to parallel performance analysis [Wolf, GRS]
- 10:30 Cachegrind [Weidendorfer, TUM]
- 12:00 (lunch)
- 13:30 *Periscope* [Gerndt, TUM]
- 15:00 Scalasca [Wylie, JSC]
- 16:30 Review of day
- 17:00 (adjourn)



Monday 1st November

- 09:00 *Vampir* [Weber, TUD-ZIH]
- 10:30 IBM HPC Toolkit [Allsopp, KSL]
- 11:00 Wrap-up discussion [Bungartz/Weinzierl, TUM]
- 11:30 (lunch)
- 13:30 KSL User Assistance [all]
 - Analysis & tuning of participants' application codes
 - Ensure your code builds and runs to completion in a reasonable time (say 15 minutes) with an appropriate dataset for initial analyses
 - Also prepare some larger/longer scalability "production" configurations (perhaps with only a few iterations/steps) to analyse if time permits
 - Consider preparing different versions (algorithms/optimizations) or systems to compare
- 17:00 (adjourn)



- Bootable Linux installation on DVD (or USB memory stick)
- Includes everything needed to try out our parallel tools on an x86-architecture notebook computer
 - GCC compiler suite (with OpenMP support), OpenMPI library
 - POINT tools: PAPI, PerfSuite, TAU
 - VI-HPS tools: Marmot, Periscope, Scalasca, VT/Vampir*
 - Other tools: BUPC, dyninst, Eclipse/PTP, PPW, TotalView*
 - * time/capability-limited evaluation licences provided for commercial products
 - Manuals/User Guides
 - Tutorial exercises and examples
- Prepared by U. Oregon Performance Research Laboratory
 - Sameer Shende