

On the ROI of Parallel Performance Optimization

Bernd Mohr (Jülich Supercomputing Centre)

EU H2020 Centre of Excellence (CoE)

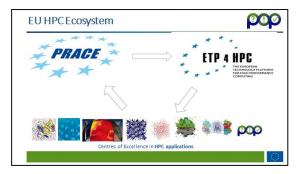


1 December 2018 – 30 November 2021

Grant Agreement No 824080

POP CoE

- A Centre of Excellence
 - On Performance Optimisation and Productivity
 - Promoting best practices in parallel programming
- Providing FREE Services
 - Precise understanding of application and system behaviour
 - Suggestion/support on how to refactor code in the most productive way
- Horizontal
 - Transversal across application areas, platforms, scales
- For (EU) academic AND industrial codes and users !







Partners



• Who?

- BSC, ES (coordinator)
- HLRS, DE
- IT4I, CZ
- JSC, DE
- NAG, UK
- RWTH Aachen, IT Center, DE
- TERATEC, FR
- UVSQ, FR

A team with

- Excellence in performance tools and tuning
- Excellence in programming models and practices
- Research and development background AND proven commitment in application to real academic and industrial use cases



3



Motivation



Why?

- Complexity of machines and codes
 - ⇒ Frequent lack of quantified understanding of actual behaviour
 ⇒ Not clear most productive direction of code refactoring
- Important to maximize efficiency (performance, power) of compute intensive applications and productivity of the development efforts

What?

- Parallel programs, mainly MPI/OpenMP
 - Although also CUDA, OpenCL, OpenACC, Python, ...



The Process ...

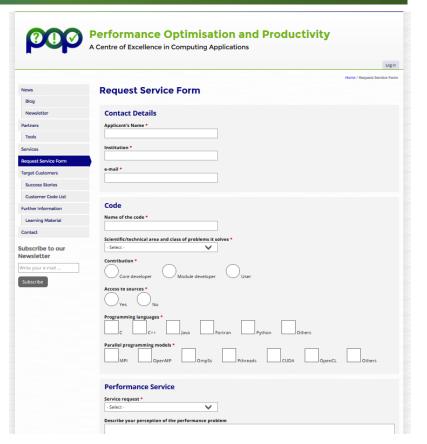


When?

December 2018 – November 2021

How?

- Apply
 - Fill in small questionnaire describing application and needs <u>https://pop-coe.eu/request-service-form</u>
 - Questions? Ask pop@bsc.es
- Selection/assignment process
- Install tools @ your production machine (local, PRACE, ...)
- Interactively: Gather data \rightarrow Analysis \rightarrow Report





FREE Services provided by the CoE



- Primary service
- Identifies performance issues of customer code (at customer site)
- If needed, identifies the root causes of the issues found and qualifies and quantifies approaches to address them (recommendations)
- Combines former Performance Audit (?) and Plan (!)
- Medium effort (1-3 months)

Proof-of-Concept (✓)

- Follow-up service
- Experiments and mock-up tests for customer codes
- Kernel extraction, parallelisation, mini-apps experiments to show effect of proposed optimisations
- Larger effort (3-6 months)



000	pap
	and the second s
The HH Concerning community of	 The action is write they firm only a MPI material are noted, and determines the same for the distribution of a matching process and process.
Manufacture Contraction	when it is reason when its 200 which makes makes as the provide its contact in 2 and of or in the odd — sample force its pro- top when of the provide its 10 minute hyperbolic and provide its minute and the contact and the force of the data.
	DETER Lander and a state of the second
See Scooling or over child at the line of the set	- atitie
Peter a far and de see	
	A A A A A A A A A A A A A A A A A A A

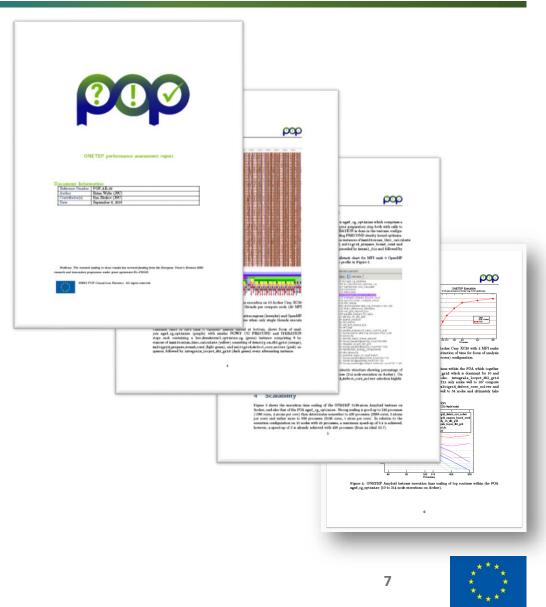




Outline of a Typical Audit Report



- Application Structure
- (If appropriate) Region of Interest
- Scalability Information
- Application Efficiency
 - E.g. time spent outside MPI
- Load Balance
 - Whether due to internal or external factors
- Serial Performance
 - Identification of poor code quality
- Communications
 - E.g. sensitivity to network performance
- Summary and Recommendations

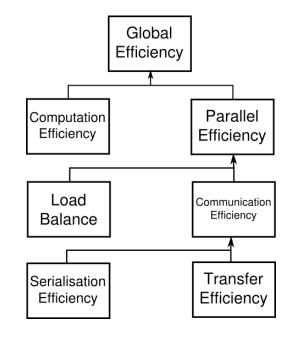


Efficiencies

• The following metrics are used in a POP Performance Audit:

- Global Efficiency (GE): GE = PE * CompE
 - Parallel Efficiency (PE): PE = LB * CommE
 - Load Balance Efficiency (LB): LB = avg(CT)/max(CT)
 - Communication Efficiency (CommE): CommE = SerE * TE
 - Serialization Efficiency (SerE):
 SerE = max (CT / TT on ideal network)
 - Transfer Efficiency (TE): TE = TT on ideal network / TT
 - (Serial) Computation Efficiency (CompE)
 - Computed out of IPC Scaling and Instruction Scaling
 - For strong scaling: ideal scaling -> efficiency of 1.0





CT = Computational time TT = Total time



Efficiencies



	2	4	8	16	
Parallel Efficiency	0.98	0.94	0.90	0.85	
Load Balance	0.99	0.97	0.91	0.92	
Serialization efficiency	0.99	0.98	0.99	0.94	
Transfer Efficiency	0.99	0.99	0.99	0.98	
Computation Efficiency	1.00	0.96	0.87	0.70	
Global efficiency	0.98	0.90	0.78	0.59	

	2	4	8	16	
IPC Scaling Efficiency	1.00	0.99	0.96	0.84	
Instruction Scaling Efficiency	1.00	0.97	0.94	0.91	
Core frequency efficiency	1.00	0.99	0.96	0.91	



Tools



Install and use already available monitoring and analysis technology

- Analysis and predictive capabilities
- Delivering insight
 - With extreme detail
 - Up to extreme scale
- Open-source toolsets
 - Extrae + Paraver
 - Score-P + Cube + Scalasca/TAU/Vampir
 - Dimemas, Extra-P
 - MAQAO

Commercial toolsets

(if available at customer site)

- Intel tools
- Cray tools
- ARM tools



Target customers



Code developers

- Assessment of detailed actual behaviour
- Suggestion of most productive directions to refactor code

• Users

- Assessment of achieved performance in specific production conditions
- Possible improvements modifying environment setup
- Evidence to interact with code provider

• Infrastructure operators

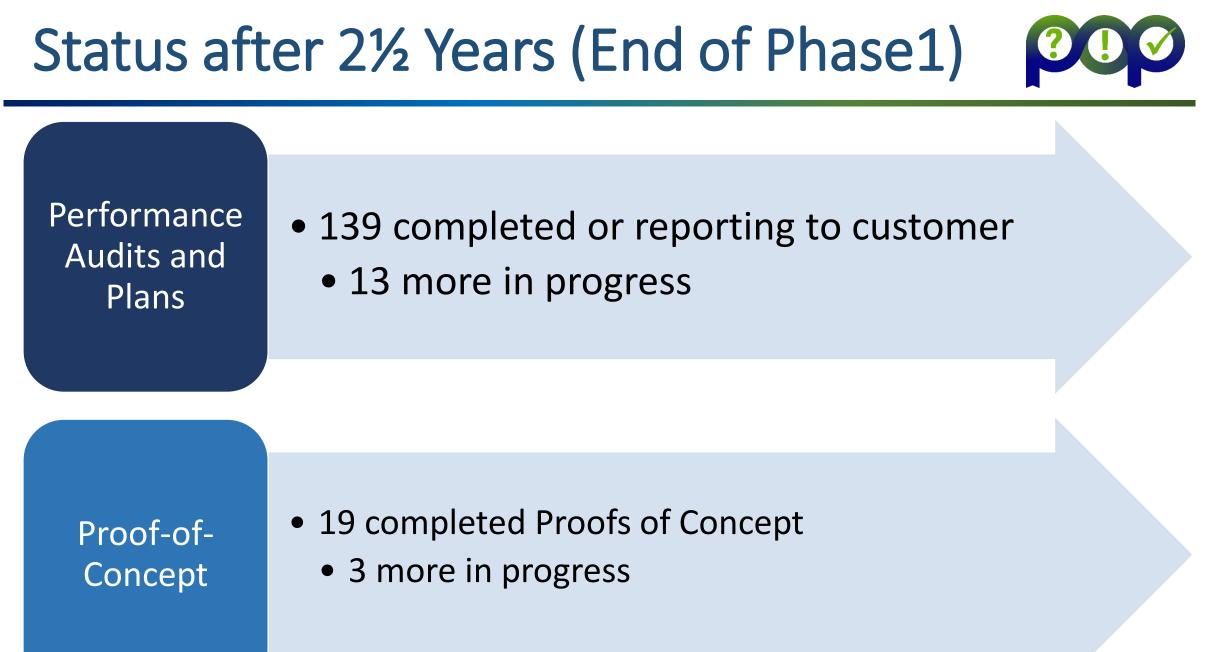
- Assessment of achieved performance in production conditions
- Possible improvements from modifying environment setup
- Information for time computer time allocation processes
- Training of support staff
- Vendors
 - Benchmarking
 - Customer support
 - System dimensioning/design





Overview of Codes Investigated







Example POP Users and Their Codes

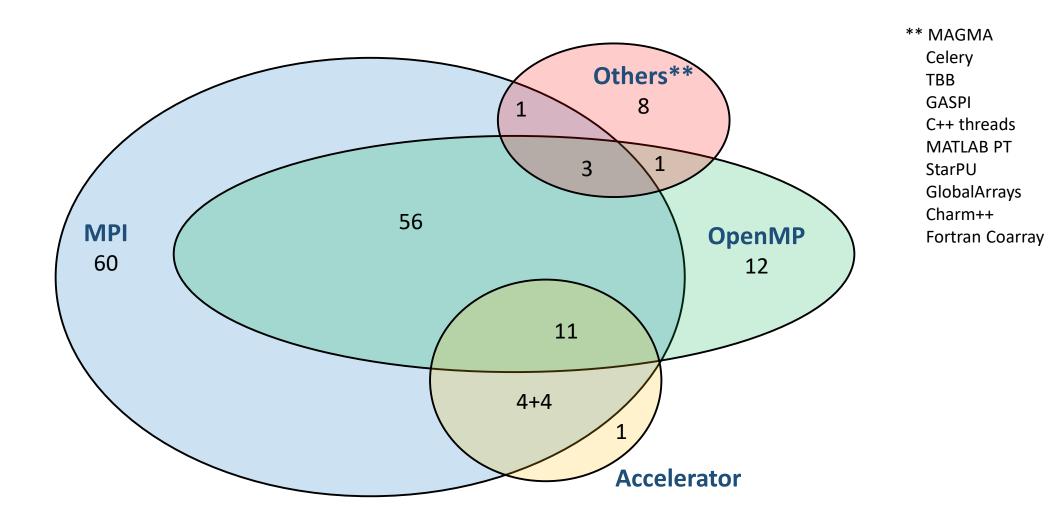


Area	Codes
Computational Fluid Dynamics	DROPS (RWTH Aachen), Nek5000 (PDC KTH), SOWFA (CENER), ParFlow (FZ-Juelich), FDS (COAC) & others
Electronic Structure Calculations	ADF, BAND, DFTB (SCM), Quantum Expresso (Cineca), FHI-AIMS (University of Barcelona), SIESTA (BSC), ONETEP (University of Warwick)
Earth Sciences	NEMO (BULL), UKCA (University of Cambridge), SHEMAT-Suite (RWTH Aachen), GITM (Cefas) & others
Finite Element Analysis	Ateles, Musubi (University of Siegen) & others
Gyrokinetic Plasma Turbulence	GYSELA (CEA), GS2 (STFC)
Materials Modelling	VAMPIRE (University of York), GraGLeS2D (RWTH Aachen), DPM (University of Luxembourg), QUIP (University of Warwick), FIDIMAG (University of Southampton), GBmoIDD (University of Durham), k-Wave (Brno University), EPW (University of Oxford) & others
Neural Networks	OpenNN (Artelnics)



Programming Models Used

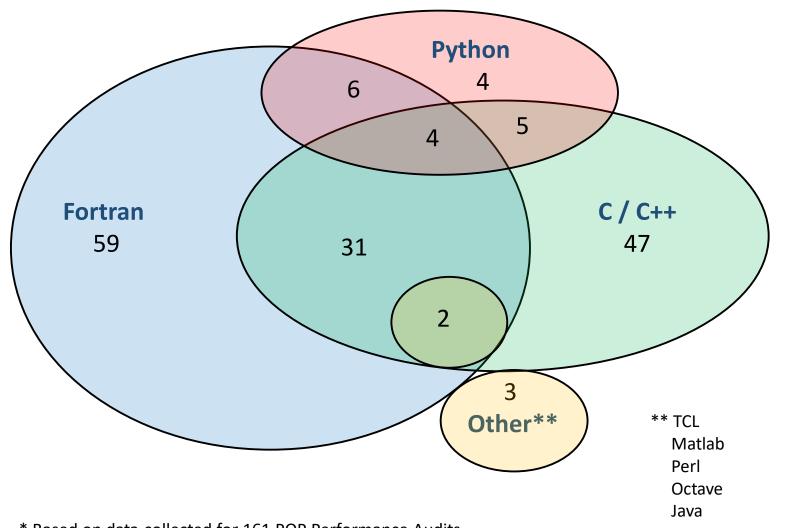






Programming Languages Used

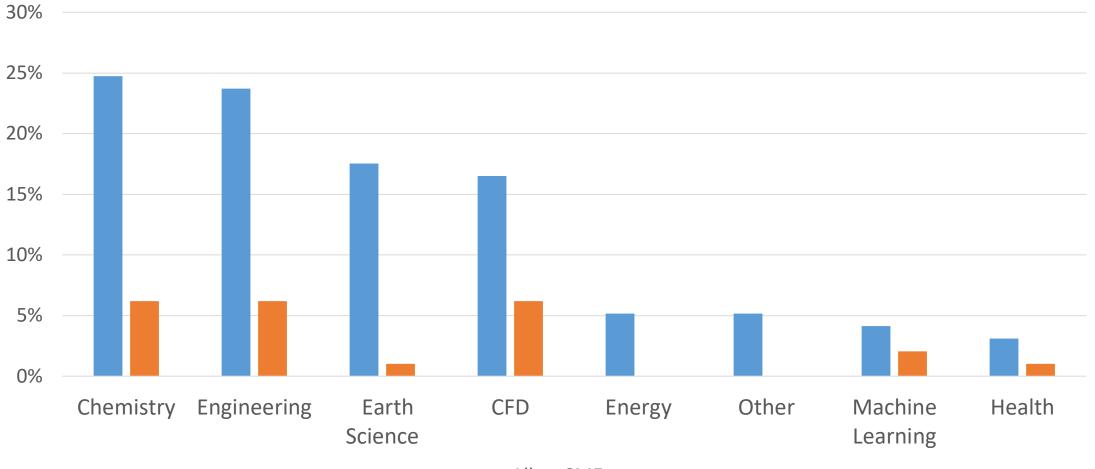




* Based on data collected for 161 POP Performance Audits



Application Sectors



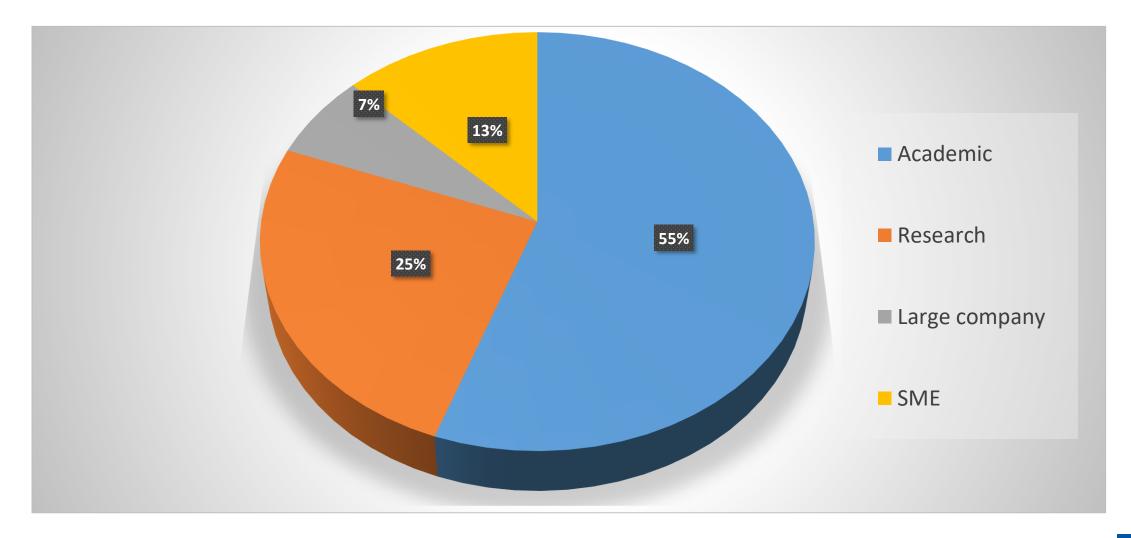
All SMEs





Customer Types









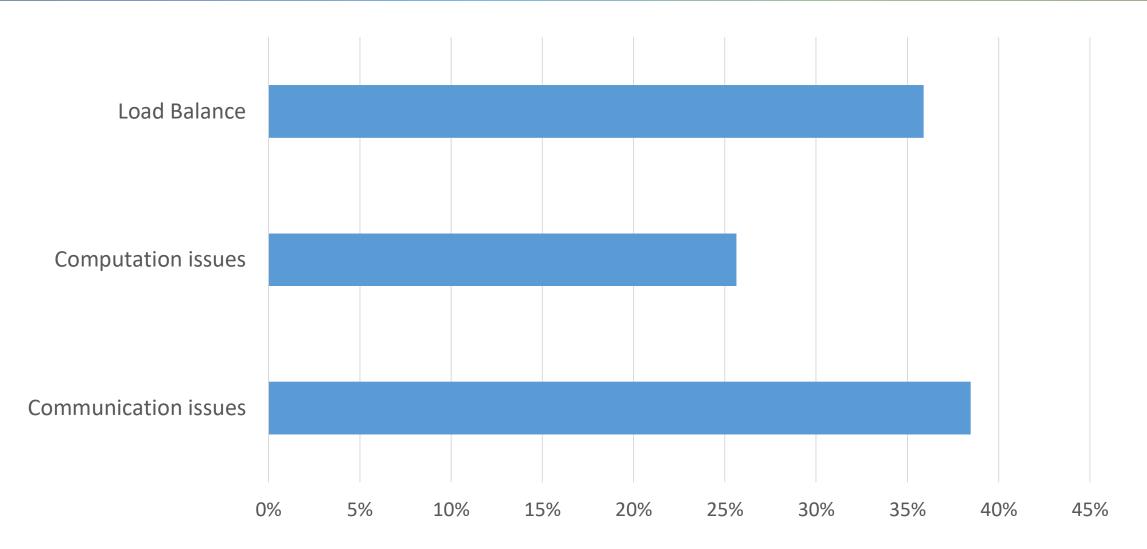
Analysis of Inefficiencies



19

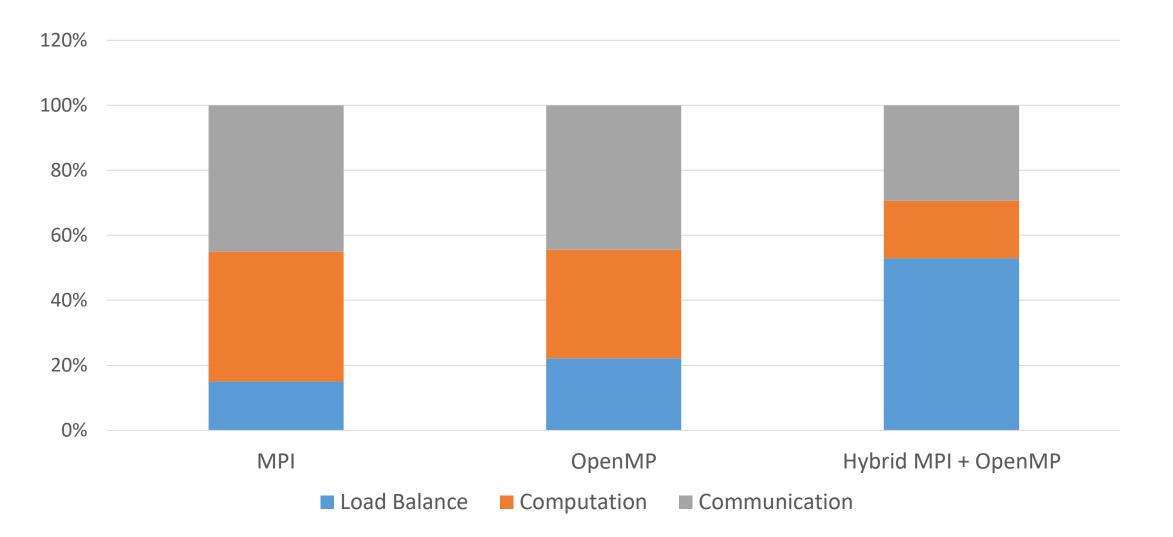
Leading Cause of Inefficiency







Inefficiency by Parallelisation





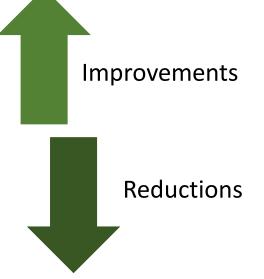
Success Stories



Some PoC Success Stories



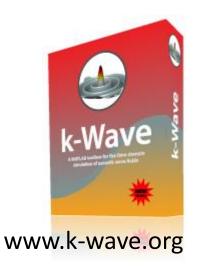
- See
 → https://pop-coe.eu/blog/tags/success-stories
 - Series Performance Improvements for SCM's ADF Modeling Suite
 - **3x Speed Improvement** for zCFD Computational Fluid Dynamics Solver
 - **25% Faster time-to-solution** for Urban Microclimate Simulations
 - **2x performance improvement** for SCM ADF code
 - Proof of Concept for BPMF leads to around **40% runtime reduction**
 - POP audit helps developers double their code performance
 - **10-fold scalability improvement** from POP services
 - POP performance study improves performance up to a factor 6
 - 1
- POP Proof-of-Concept study leads to nearly 50% higher performance
- POP Proof-of-Concept study leads to **10X performance improvement** for customer





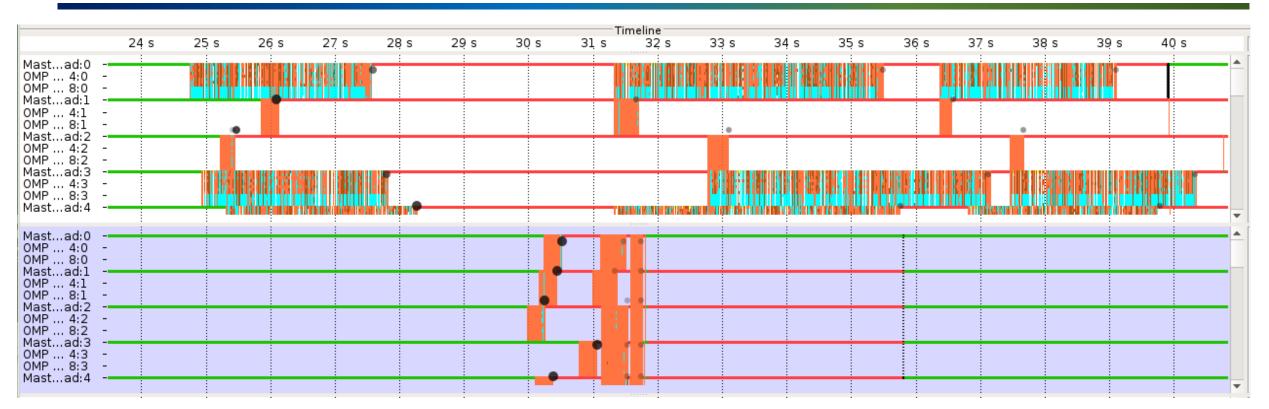
k-Wave – Brno Uni. of Technology

- Toolbox for time domain acoustic and ultrasound simulations in complex and tissue-realistic media
- C++ code parallelised with Hybrid MPI and OpenMP (+ CUDA)
- Executed on Salomon Intel Xeon compute nodes
- Key audit findings:
 - 3D domain decomposition suffered from major load imbalance : exterior MPI processes with fewer grid cells took much longer than interior
 - OpenMP-parallelised FFTs were much less efficient for grid sizes of exterior, requiring many more small and poorly-balanced parallel loops
- Using a periodic domain with identical halo zones for each MPI rank reduced overall runtime by a factor of 2





k-Wave – Brno Uni. of Technology

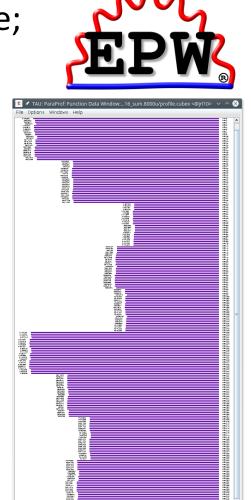


- Comparison time-line before (white) and after (purple) balancing, showing exterior MPI ranks (0,3) and interior MPI ranks (1,2)
 - User code in green, MPI synchronization in red, OpenMP synchronization in cyan



EPW – University of Oxford

- Electron-Phonon Wannier (EPW) materials science DFT code;
- part of the Quantum ESPRESSO suite
- Fortran code parallelised with MPI
- Audit of unreleased development version of code
- Executed on ARCHER Cray XC30 (24 MPI ranks per node)
- Key audit findings:
 - Poor load balance from excessive computation identified
 - (addressed in separate POP Performance Plan)
 - Large variations in runtime, likely caused by IO
 - Final stage spends a great deal of time writing output to disk
- Report used for successful PRACE resource allocation



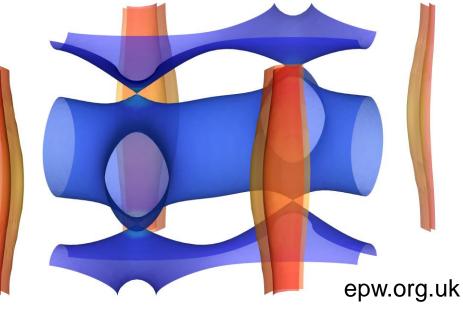




EPW – University of Oxford



- Original code had all MPI ranks writing the result to disk at the end
- POP PoC modified this to have only one rank do output
- On 480 MPI ranks, time taken to write results fell from over 7 hours to 56 seconds: 450-fold speed-up!
- Combined with previous improvements, enabled EPW simulations to scale to previously impractical 1920 MPI ranks
- 86% global efficiency with 960 MPI ranks





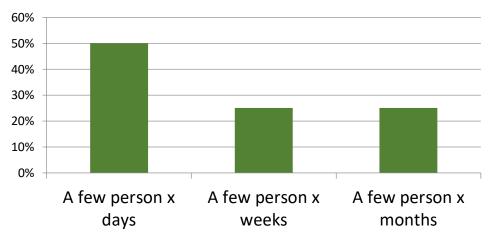
(Eight) Customers Success Feedback



What is the observed performance gain after implementing recommendations?

25%
25%
20% overall, 50% for the given module
50-75% (case dependent)
12%
Up to 62 %, depending on the use case.
6 - 47 % depending on the test case.
15%

How much effort was necessary?



80% 70% 60% 50% 40% 30% 20% 10% 0% Possibility to Possibility to Possibility to Other (please Only Better performance scalability better exploit run on a treat larger specify) slower problems gain new platform architectures (handling the (mixing multisame problem and many-

size)

What are the main results?



core servers)



Summary & Conclusion



29

Customer Acquisition



Interactions with Leads

- 86% of users needed multiple interactions before signing up
 - Users with only 1 interaction referred by existing users
- Average number of interactions to sign up = 3.2
- Maximum number of interactions to sign up = 11

Conversions

- Over 1300 leads contacted throughout the project
- Conversion rate of 10.8% from leads to user
- Only 17 signed up without direct contact from POP



Costumer Feedback



Performance Audits (73 customers)	 About 90% very satisfied or satisfied with service About half of the customers signed-up for a follow-up service
Performance Plans (11 customers)	 About 90% very satisfied or satisfied with service All customers thought suggestions were precise and clear and 70% plan to implement the suggested code modifications About 2/3 plan to do use the POP services again
Proof-of-Concepts (8 customers)	 All customers very satisfied or satisfied with this service About 80% plan to implement further code modifications or complete the work of the POP experts

* Based on data collected in 92 customer satisfaction questionnaires and 52 phone interviews with customers



ROI Examples



Application Savings after POP Proof-of-Concept

- POP PoC resulted in 72% faster-time-to-solution
- Production runs on ARCHER (UK national academic supercomputer)
- Improved code saves €15.58 per run
- Yearly savings of around €56,000 (from monthly usage data)

Application Savings after POP Performance Plan

- Cost for customer implementing POP recommendations: €2,000
- Achieved improvement of 62%
- €20,000 yearly operating cost
- Resulted in yearly saving of €12,400 in compute costs ⇒ ROI of 620%



Summary & Conclusion (I)



- POP CoE Phase 1 finished in March 2018 after 2½ years
 - Successfully demonstrated expertise and impact
 - 152 Audits + Perf Plans / 22 Proof-of-Concept / 21 requests cancelled
 - 158 closed / 16 in progress
 - Intensive dissemination via website, blog articles, tweets, newsletter, ...

⇒ Expected more interest from industry / SME / ISVs

- POP CoE Phase 2 restarted in December 2018 for 3 more years
 - New Service Structure (Performance Assessment combines Audit and Plan)
 - New Project Partners (IT4I, UVSQ)
 - New Co-design Data Repository
 - Extension of Efficiency Model: Vectorization, I/O, GPUs, ...



Summary & Conclusion (II)



- Issues identified
 - FREE (Money) ≠ FREE (Efforts, Time)
 - To much(?) customer effort (providing code, input, measurements?, feedback)
 - Desire to serve more industrial customers / SMEs, BUT
 - Resistance for allowing us to publish their results / success stories
 - Almost every time require NDA agreements
 - Sustainability
 - Real costs audit (EUR 16K-18K) >> Price customer would pay (5K-7K)





Dissemination and Contact



35

Website – www.pop-coe.eu



- POP User Portal
- Access to all public information and services



Blog – <u>https://pop-coe.eu/blog</u>



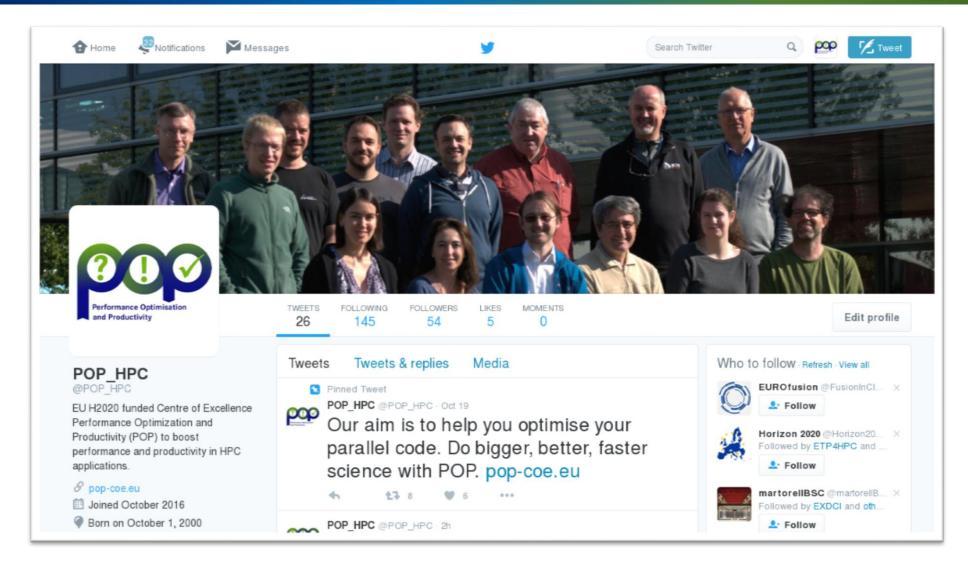
- Typically 2 new articles per month
- Easy filtering via Tags, e.g
 - Success Stories
 - Events
 - Webinars
 - ...
- RSS feed
 - https://pop-coe.eu/blog/rss





Follow us on Twitter @POP_HPC

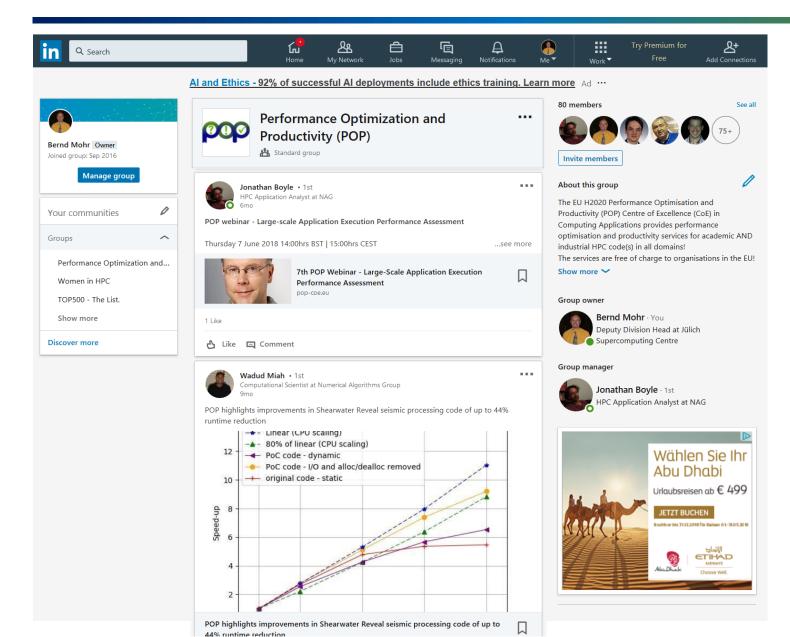






LinkedIn Group





- Important announcements
- Serves also as user forum

39



Quarterly Email Newsletter



- Subscribe on POP website
- Newsletter archive: https://pop-coe.eu/news/newsletter



POP Partner Profiles

Group at Ruith Aacher

The KHO COL Investor Approximation The Benefities Representation of the Microsoft Computing Center Budger (M-R), unlish Representation Center (JAC), Numerical Approximation Group (M-G), NITTIN Advantant TEANTECC.

The Numerical Algorithms Group (NAG) NG product netral offset sounding of the activities sounding seems to sound an industry indicate from the sound activities of the activ

behaves the unsensities of Birmingham, Laess. Manchester: Notifingham and Orders, and wrat is now second as the Rutherbox Appartm Laessatory in 1371 valdesetspective NAG Library which has subsequently accore the angest commercially available collection of high quarky instimutations and satisfactural approxim-



Averged to not in the HPC feet, NeQ protects as range of annotes to HPC organisations from an implement summaria computation activities. These mouses the same mission is not annote proteiner and writing that obser applications remain valid and optimal to the lated processions and patterns. Additionally NAG is same initiated to trag observe external-produced invariant libraries.

The HPC Group at RWTH Aachen



Alth-denses +CC course via approximation to access the date of the Alth-dense +CC course of the Alth-dense approximation of t

being there year in the second of the second

POP's Performance Analysis Tools

Two memoers of the POP partnership owerko poverful performance analysis boos which are used edensively in our work.

The Barcelona Supercomputer Center (BSC) Performance Tools

The performance analysis tools developed at BBC protoe a detailed analysis that allows understanding of an application's behaviour as veri as identifying performance orthoat leaves. They provide marget not only into the application them out also into the underlying system.

The cost on Anterex, a trade-searce profession and web grant famility is export and extra information. Fasters graneses to runn exacultations termina, and the graneses being provide to the grant family the sector of the sector and the sector and the graneses being information. These to exponentity we also searce extra family conditions that previous devices and the grant family sector and the graneses being sector and the sector and the grant family sector and the grant fam

Purtnemore. Fursherr contains anarytic modules, for example the clustering too for semi-submatic detection of the application envolues, and the tracking too to detection of the application envolues, and the tracking too to detection of the application envolues.

In aptition, the Dmenuse simulator above a fast evaluation of unainfrequenties for message passing approaching, for example to understand the Seretta of moving to machine with a faster refurer, or the improvements citative of Fan application valo wefter plannad.

The performance data collection is one with Schaie Schaie International parallel ruther-exercisioners (API, OperAP Orgale Perseau, CUCA, OperO, Enricitios, and support and any improveming any page 42. C + A frame. Perform Anal, The later exercision (API, OperAP) and the C for Budden Public Schair (March Schaie), and early Arodic O in racipations the persoan resonance access as to acce specific constraints and the memory for persoance in program.

The performance book from ESC are freely available as Open Source software. You can find none information at our resolute the test are an an an advantage of a software exception of the software except

new blog post - tox improvement for openime code

OpenLIP activate obtained a 10-performance improvement on 125 core SMP instakers after a POP performance analysis, improvement plan and proof-of-concept React the blog post terms

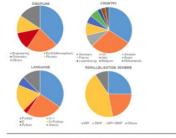
Breakdown of our users' codes

In this article we take a curck took at the type of software profiled since the start of the POP project to demonstrate the preach of our

Fengs unusingly, not applications do trais aler this tractional despites to ref. Lis engenering and Lisenspites exercise, serving are phototranspitorent granupas auto) pro cines an indusen moly tractionary new company. In 6 Yens, C. C. H. Housey more C. C. L. Housey more C. C. Housey and the application of the state of the serving are arranging tracking to the form units and an energing transpitor to the state of the serving are arranging tracking to exclusion. The serving are arranging tracking to exclusion of the serving transpitor to exclusion of the serving transpitor transmission of the serving transmission of the serving transpitor transmission of the serving tra

A brailedow of parallel nethoology shows a large adoption of hjöric parallelism using MPI + OpenMP which makes profiling nore challenging. The "Ober's catego norules uses threads. Phreads. Tote & PGAB.

Vertible very interested in hearing from code beveropers in other disciplines and from other EU countries.



Meet POP CoE partners at some upcoming events

POP (a) ISC 2016 Cree to or POP 6/F at SC 2016 of use 11 or rest or egent at he research boths of SSC (#22), NAC (#518) or USCH-RRS (#10).

POP © Teratec Forum

POP represents by BSC and NAS will be at the "European Research Cafe" in the <u>Terrer Forum</u> on June 28 and 26, 2016 in France POP & HEPCS 2016

POP represented by NAG, will present a poster at the <u>HPCB 2016 configurous</u> on July 18 to 22, 2016 in Austria.







Webinars / YouTube

- Or see our
 channel <u>youtube.com/POPHPC</u>
- Already available:
 - How to Improve the Performance of Parallel Codes
 - Using OpenMP Tasking
 - Parallel I/O Profiling Using Darshan
 - Getting Performance from OpenMP Programs on NUMA Architectures
 - Understand the Performance of your Application with just Three Numbers
 - The impact of sequential performance on parallel codes
 - Large scale Application Execution Performance Assessment
 - POP Case Study: 3x Speed Improvement for Zenotech's zCFD Solver
 - Exascale Matrix Factorization: Using HPC and ML for Drug Discovery
 - Software for Linear Algebra Targeting Exascales (SLATE) Project
 - Implementing I/O Best Practices to Improve System Performance with Ellexus









Performance Optimisation and Productivity A Centre of Excellence in HPC

Contact: https://www.pop-coe.eu mailto:pop@bsc.es ≥ @POP_HPC ▷ youtube.com/POPHPC





This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 676553 and 824080.