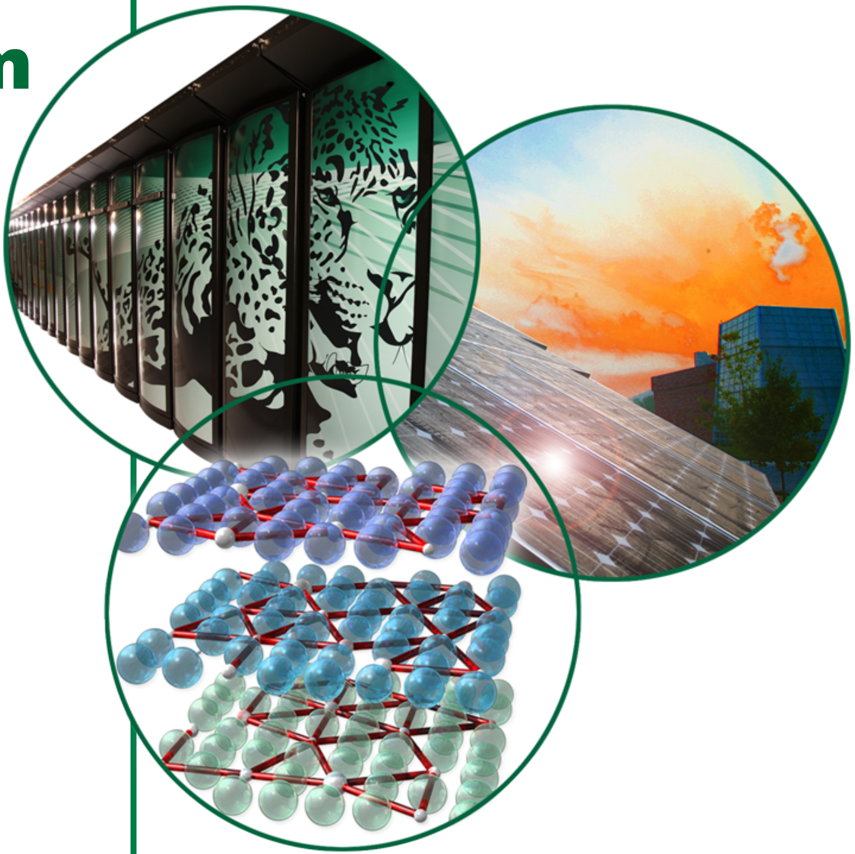
Runtime Tracing of the Community Earth System Model: Feasibility Study and Benefits

ICCS'12 Workshop -
Tools for Program Development and Analysis in Computational Science



Jens Domke, JICS, ORNL

June 05, 2012

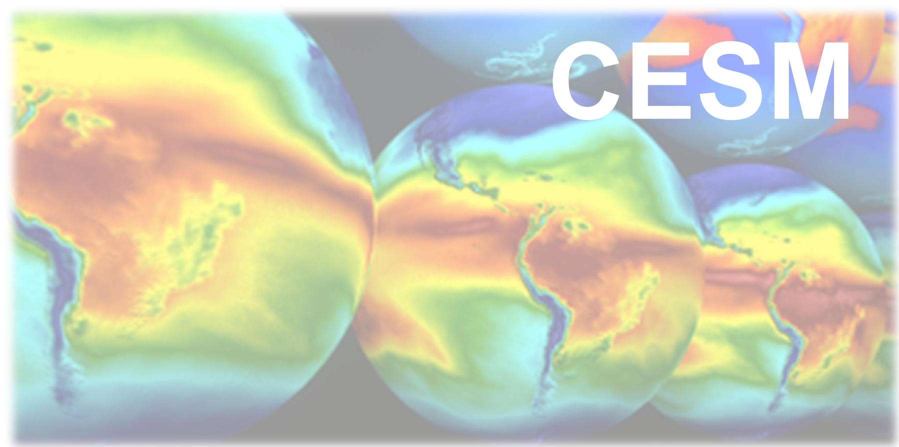
U.S. DEPARTMENT OF ENERGY

OAK RIDGE NATIONAL LABORATORY

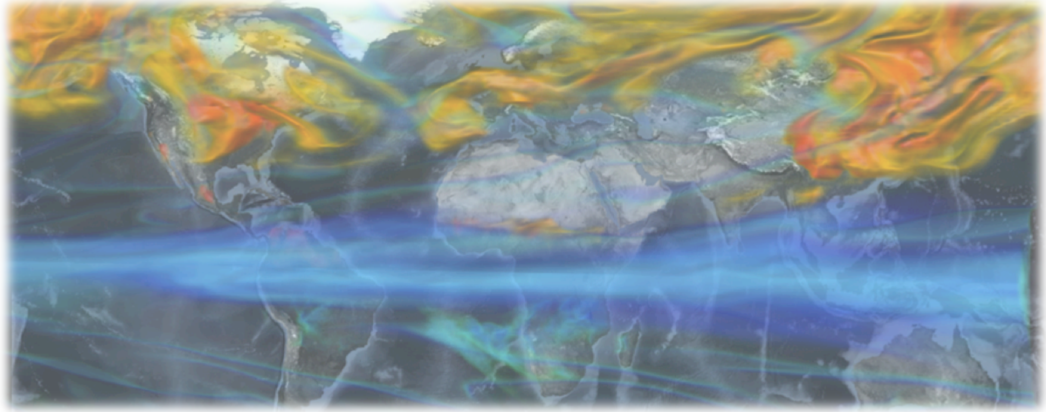
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

Agenda

1. Introduction
   - Community Earth System Model
   - Performance analysis toolset: Vampir
   - Motivation
2. Tracing of CESM
3. Outcome of the tracing
4. Summary & Conclusion



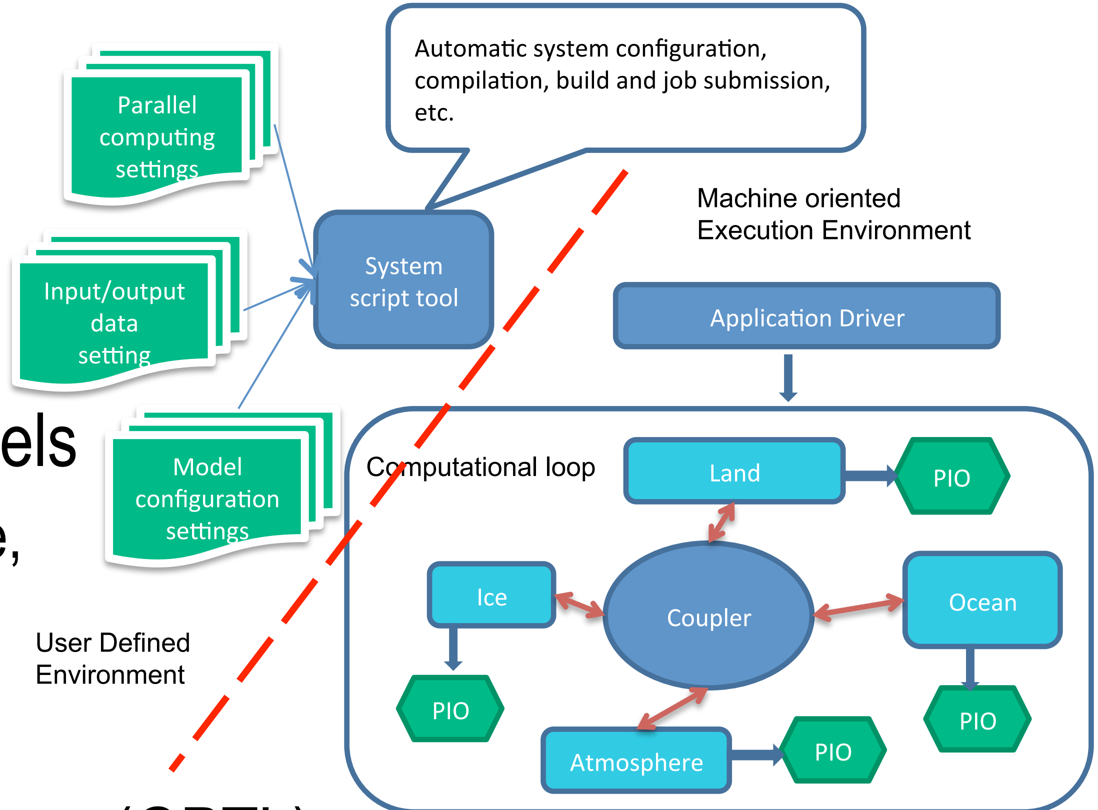
1.1 Community Earth System Model



- One of US's leading earth system modeling frameworks maintained by NCAR
- Early version where developed in the 1980s (Community Climate Model)
- Steady improvements and renaming over last decades
- Intergovernmental Panel on Climate Change (IPCC) uses CESM (among others) for climate reports/forecasts

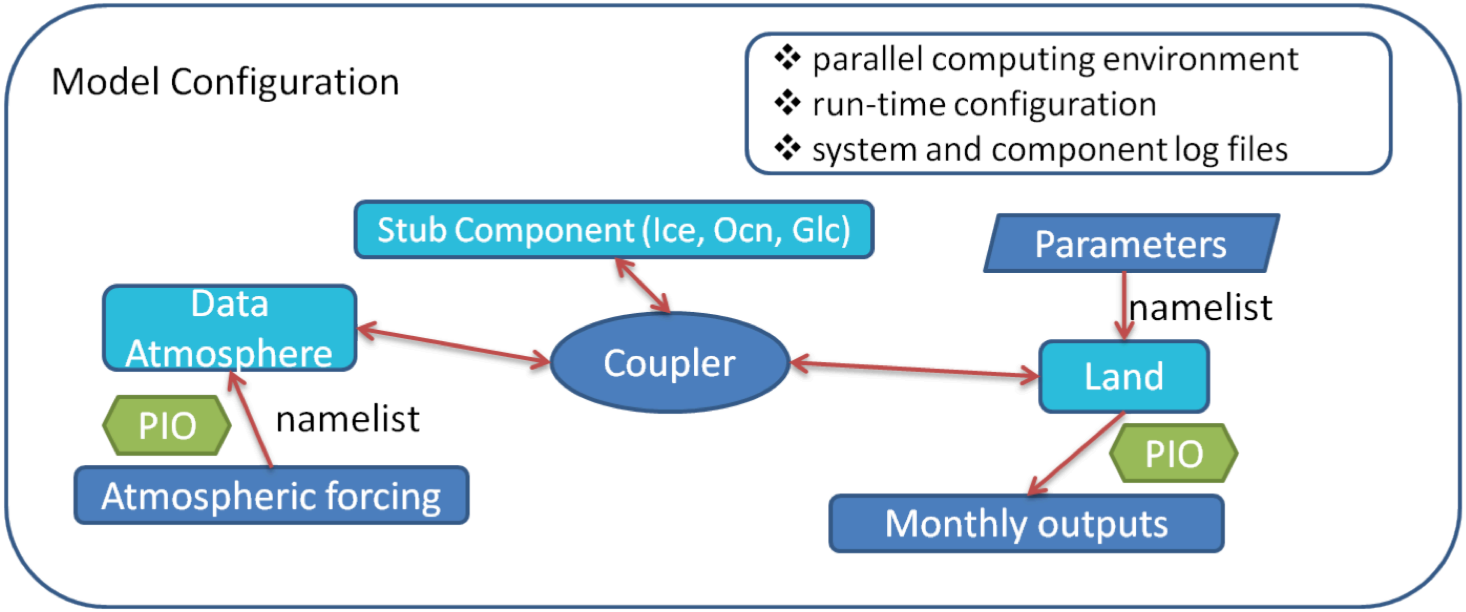
1.1 Community Earth System Model

- Build/configuration system uses C-shell scripts
  - Compilation; configuration; job submission
- Five community model components and data models
  - Atmosphere, ocean, sea ice, land, and land ice sheet
- Coupler and parallel I/O
- General purpose timing library (GPTL)
  - For profiling and access to PAPI counters



Oak Ridge National Laboratory

1.1 Community Earth System Model

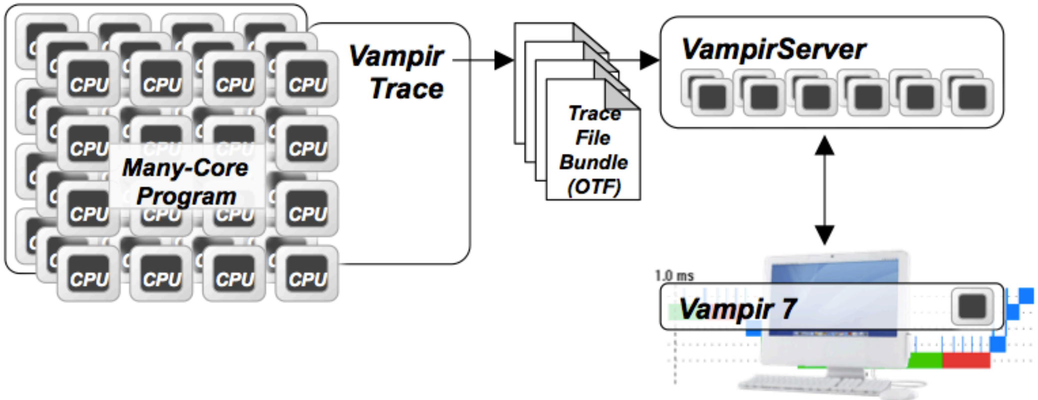


Configuration for simulations on a XT5 (Jaguar, at ORNL)

- Offline global community land model simulation
  - Data atmosphere model (DATM) and active Community Land Model (CLM4)
  - CLM4 with activated CLM-CN (carbon and nitrogen cycle simulation)
  - Stub models for ocean, ice, and glacier

1.2 VampirTrace & Vampir

- **VampirTrace**
  - Application instrumentation
  - Via compiler wrapper, library wrapper and/ or third-party software
  - Measurement
  - Event collection (functions calls, MPI, OpenMP, performance counter, memory usage, I/O, GPU)
- **Vampir (Client and Server)**
  - Trace visualization software
  - Show dynamic run-time behavior graphically
  - Provide statistics and performance metrics
  - Interactive browsing, zooming, selecting capabilities
- **Performance analysis and identification of bottlenecks, e.g.**
  - Most time consuming functions
  - Inefficient communication patterns
  - Load imbalances
  - I/O bottlenecks



1.3 Motivation

- General questions:
  - Can VampirTrace generate traces for CESM? (Feasibility study)
  - Will those traces reveal more information, compared to the integrated GPTL? (Benefits)
  - What can we learn from
    - MPI and I/O analysis
    - PAPI counters

    for further developments and simulations?

Agenda

1. Introduction
   - Community Earth System Model
   - Performance analysis toolset: Vampir
2. **Tracing of CESM**
3. Outcome of the tracing
4. Summary & Conclusion

2. VampirTrace Configuration

- Macros.<casename>
  - FC := vtf90 -vt:f90 ftn -vt:mpi -vt:inst tauinst -vt:tau -f -vt:tau tau.selective -vt:cpp fpp -vt:preprocess
  - CC := vtcc -vt:cc cc -vt:mpi -vt:inst tauinst -vt:tau -f -vt:tau tau.selective
- TAU instrumentor ➔ filter functions w/ short duration
- ‘-vt:tau -f -vt:tau tau.selective’ ➔ fix for build system
- ‘-vt:cpp fpp -vt:preprocess’ ➔ TAU problem w/ macros

2. VampirTrace Configuration

- File tau.selective:
  - Exclude list for functions with >5.000 calls per process (gathered w/ profiling mode: setenv VT_MODE ‘STAT’)
  - Exclude GPTL functions
- Problems w/ PGI Fortran preprocessor
  - fpp – bash script to run pgf90 w/ correct flags and redirect output
- File env_mach_specific
  - module load vampirtrace tau papi
  - setenv VT_IOTRACE ’yes’
  - setenv VT_METRICS ’PAPI_FP_OPS:PAPI_L2_TCM:PAPI_L2_DCA’
  - setenv VT_BUFFER_SIZE 512M

Agenda

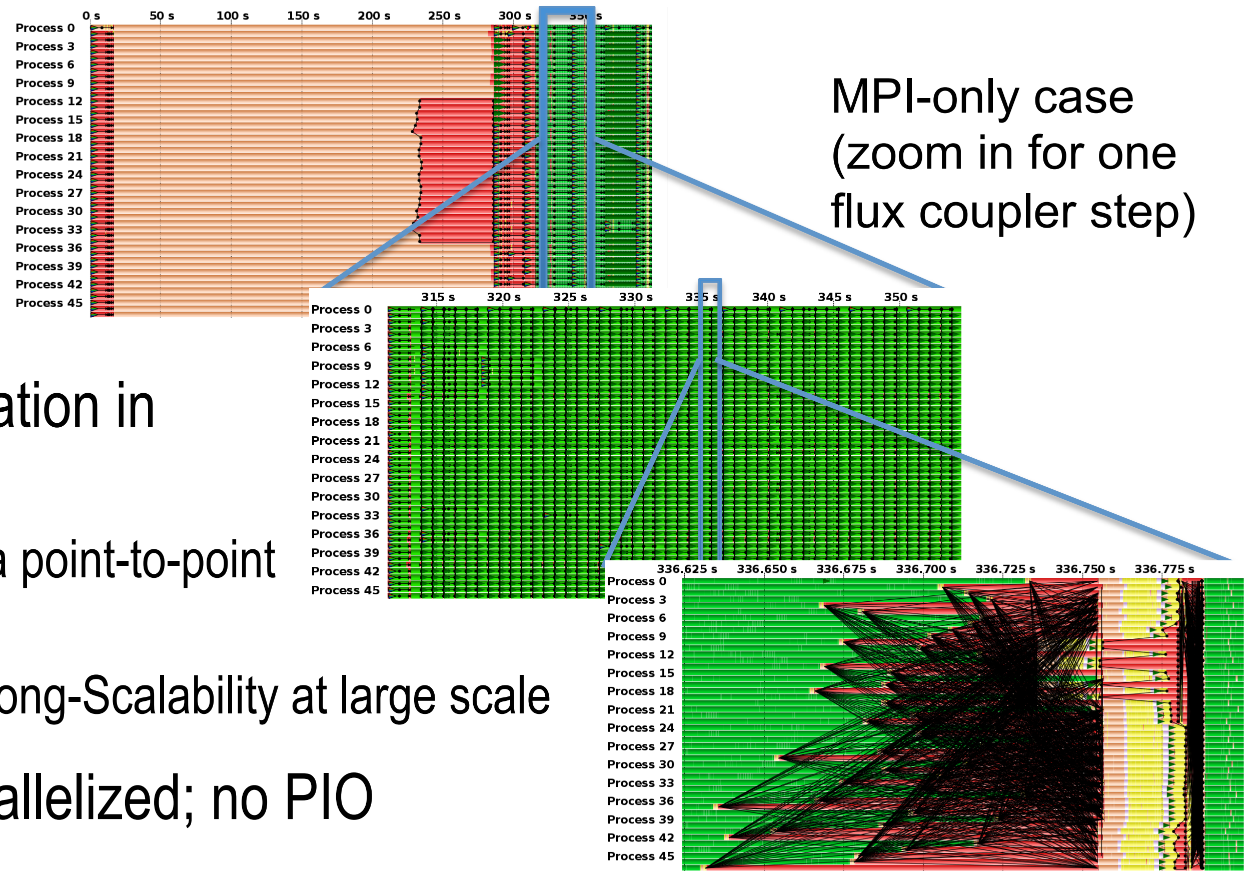
1. Introduction
   - Community Earth System Model
   - Performance analysis toolset: Vampir
2. Tracing of CESM
3. **Outcome of the tracing**
4. Summary & Conclusion

3. Simulation configuration

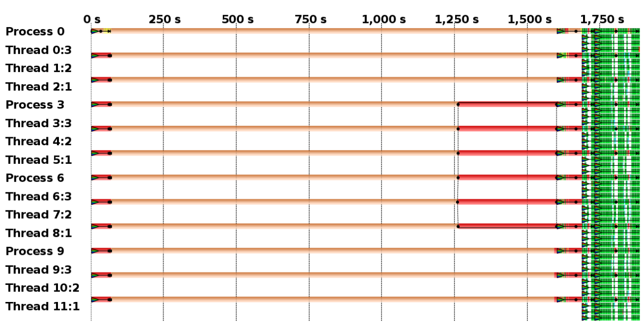
- Short-term simulation
  - 2 days of simulated climate w/o intermediate restart files
  - 48 cores (4 nodes) on a XT5
    - 48 MPI processes
    - 12 MPI processes + 4 OpenMP threads
  - Functions, I/O events, PAPI counters, MPI, OpenMP tracing
- Long-term simulation
  - One year simulation in four segments; 3 months each (using restart file of previous segment)
  - 240 MPI processes on 240 cores (20 nodes); no OpenMP
  - Only PAPI counters and MPI tracing

3.1 Tracing the short-term simulation

- Flux coupler runs every 30 min of simulated time
- Heavy global communication in flux coupler
  - Small messages send via point-to-point communication
  - ➔ One reason for poor Strong-Scalability at large scale
- DATM: not OpenMP-parallelized; no PIO



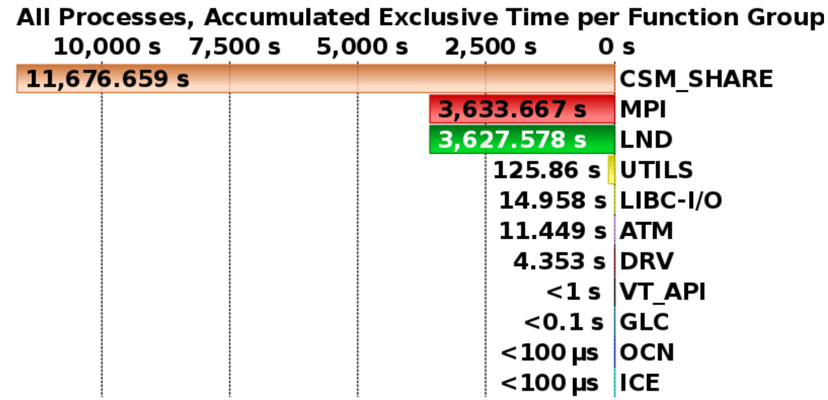
MPI-only case (zoom in for one flux coupler step)

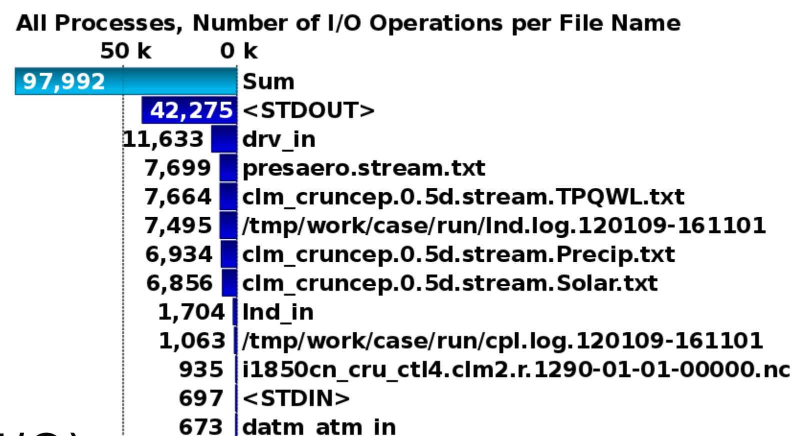


MPI+OpenMP case

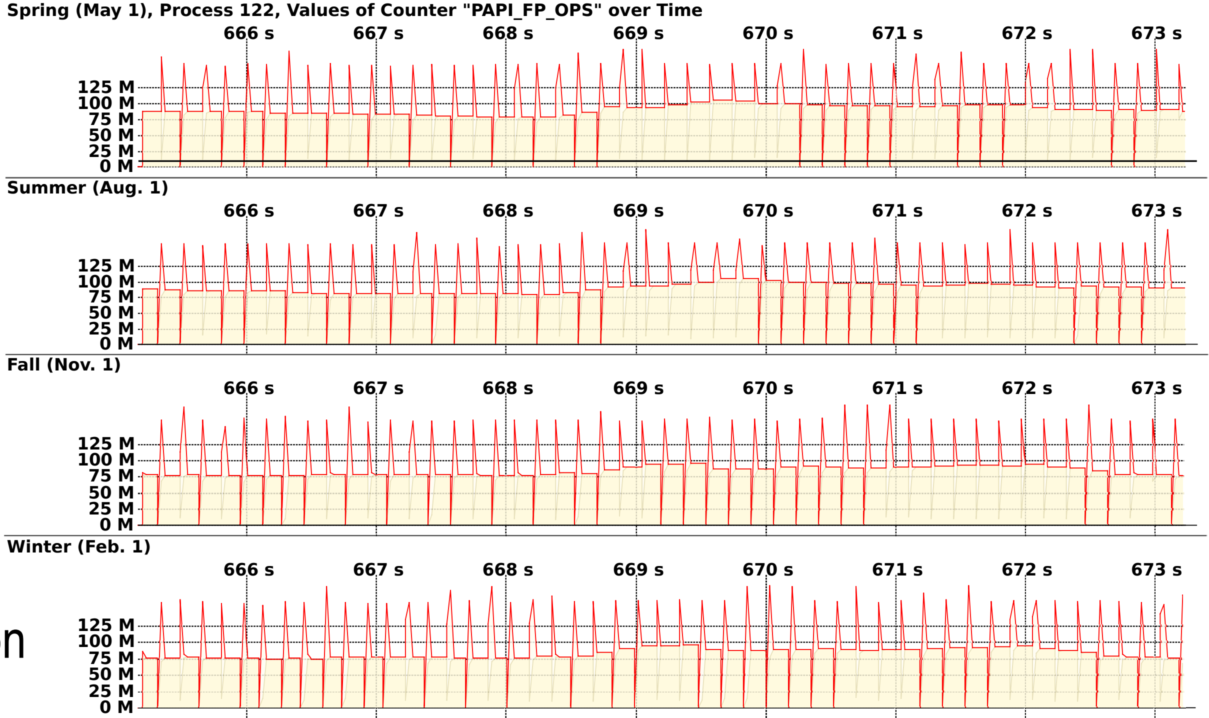
3.1 Tracing the short-term simulation

- CSM_SHARE: DATM is interpolating climate forcings
- High percentage of MPI
  - Mostly related to imbalance in DATM and MPI_Allreduce
  - Only ≈ 15% MPI within land model
- Most I/O is produced by writing timing information to stdout; rest is reading configuration files (drv, lnd, datm, …) and writing log files
- BUT: I/O is not a bottleneck (see LIBC-I/O)





3.2 Tracing the long-term simulation



Process with deciduous forest, 24 h time frame, (midnight to midnight)

- Computational intensity varies during the 24 h
  - Low flop/s counter at night
  - High counter in the afternoon
- Computational intensity of ≈ 76 Mflop/s–96 Mflop/s in winter and fall
- Spring and summer: ≈ 80 Mflop/s–106 Mflop/s
- Reason: strong relationship between land characteristics (e.g. photosynthesis) and climate forcings (like solar radiation, temperature, …)

Agenda

1. Introduction
   – Community Earth System Model
   – Performance analysis toolset: Vampir
2. Tracing of CESM
3. Outcome of the tracing
4. **Summary & Conclusion**

4. Summery & Conclusion

- CESM is traceable with low overhead
- VT/Vampir+TAU reveal more information without implementation overhead compared to GPTL
  - Partial automatic data analysis and visual processing
  - But some manual tuning is needed
- I/O operations could be excluded as possible bottleneck
- Heavy global MPI communication in flux coupler
  - Contributes to poor Strong Scalability above 768 cores

4. Summery & Conclusion

- Fine-grained performance analysis with PAPI counters
  - Variance of flop/s counter coupled to the altitude of the sun
  - Seasonal changes in computational intensity via flop/s counter visible
  - Potential to identify short-term climate extremes (like spring freeze or fire); not possible with monthly output
- Future improvements (potential was seen in the traces):
  - Dynamic load balancing during the simulation
  - OpenMP-parallelized implementation of DATM
  - Reduced overhead of flux coupler and timing management utilities

Acknowledgement

- Support by Vampir Team of the Center for Information Services and High Performance Computing (ZIH), Technische Universität Dresden
- Funding from Terrestrial Ecosystem Sciences (TES) Program and from Climate Sciences for Sustainable Energy Future (CSSEF) Program
- Access to resources of Oak Ridge Leadership Computing Facility (OLCF's Jaguar XT5 supercomputer), Oak Ridge National Laboratory