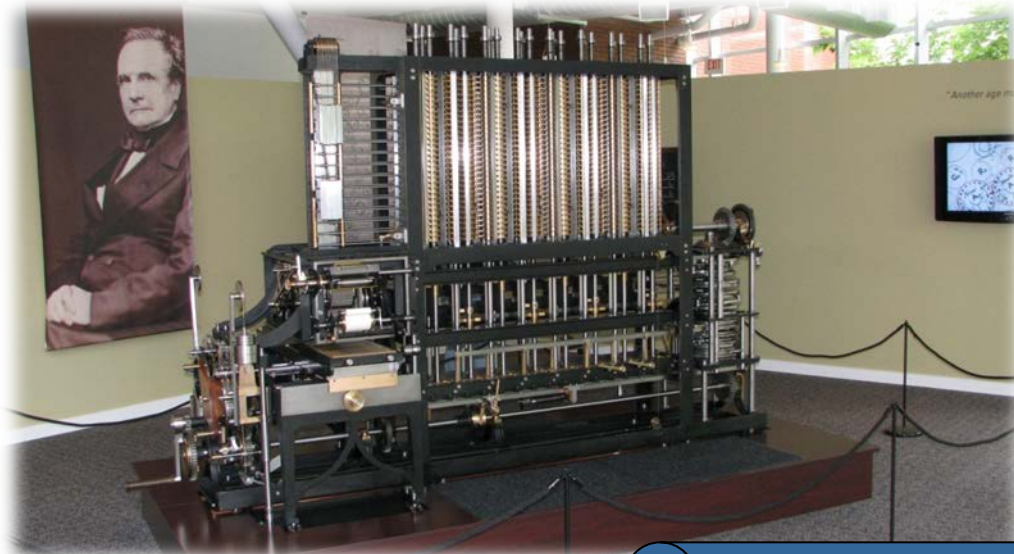




# Introduction to Parallel Performance Engineering

Brian Wylie  
Jülich Supercomputing Centre

(with content used with permission from tutorials  
by Bernd Mohr/JSC and Luiz DeRose/Cray)

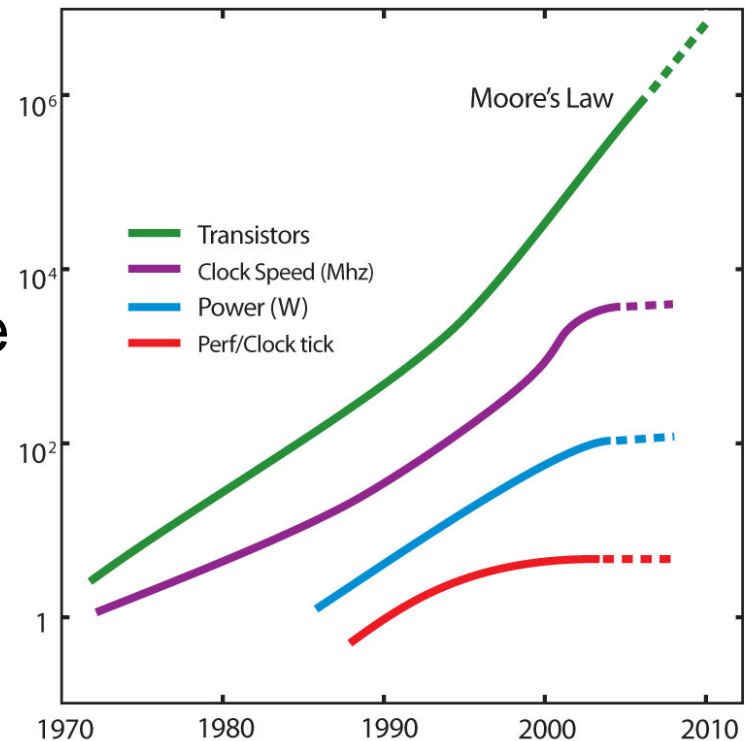


Difference Engine

“The most constant difficulty in contriving the engine has arisen from the desire to reduce the time in which the calculations were executed to the shortest which is possible.”

Charles Babbage  
1791 – 1871

- Moore's law is still in charge, but
  - Clock rates no longer increase
  - Performance gains only through increased parallelism
- Optimizations of applications more difficult
  - Increasing application complexity
    - Multi-physics
    - Multi-scale
  - Increasing machine complexity
    - Hierarchical networks / memory
    - More CPUs / multi-core



👉 Every doubling of scale reveals a new bottleneck!

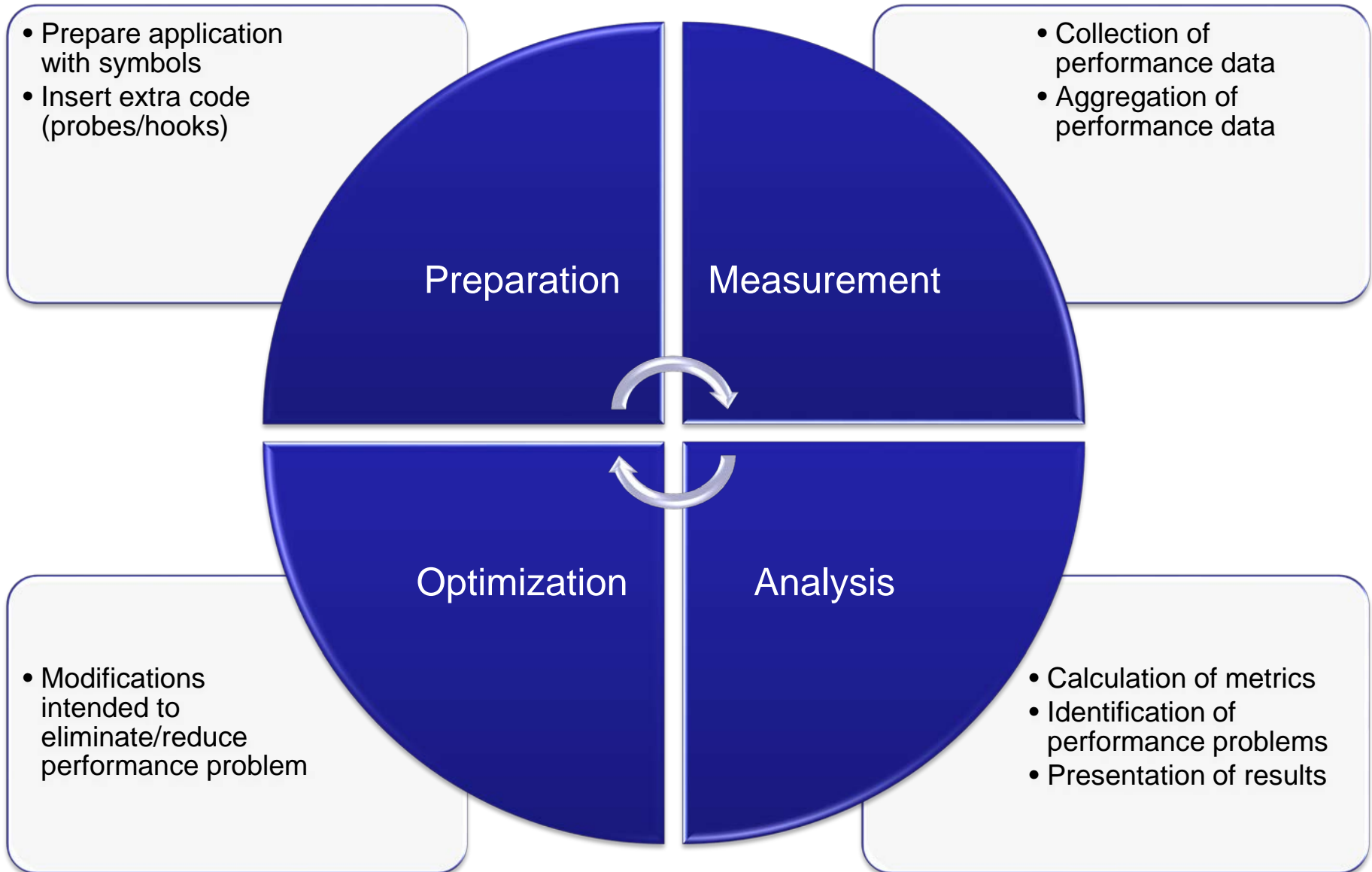
- **“Sequential”** performance factors
  - Computation
    - ☞ Choose right algorithm, use optimizing compiler
  - Cache and memory
    - ☞ Tough! Only limited tool support, hope compiler gets it right
  - Input / output
    - ☞ Often not given enough attention
  
- **“Parallel”** performance factors
  - Partitioning / decomposition
  - Communication (i.e., message passing)
  - Multithreading
  - Synchronization / locking
    - ☞ More or less understood, good tool support

- Successful engineering is a combination of
  - The right algorithms and libraries
  - Compiler flags and directives
  - Thinking !!!
- Measurement is better than guessing
  - To determine performance bottlenecks
  - To compare alternatives
  - To validate tuning decisions and optimizations
    - ☞ After each step!

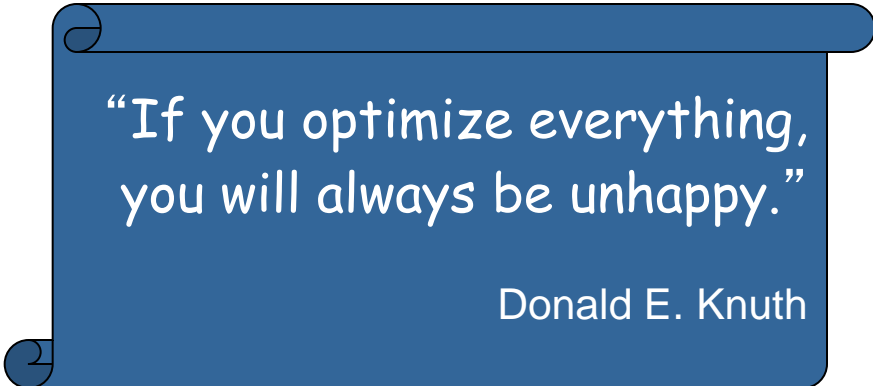
“We should forget about small efficiencies,  
say 97% of the time: premature optimization  
is the root of all evil.”

Charles A. R. Hoare

- It's easier to optimize a slow correct program than to debug a fast incorrect one
  - ☞ *Nobody cares how fast you can compute a wrong answer...*



- Programs typically spend 80% of their time in 20% of the code
- Programmers typically spend 20% of their effort to get 80% of the total speedup possible for the application
  - ☞ *Know when to stop!*
- Don't optimize what does not matter
  - ☞ *Make the common case fast!*



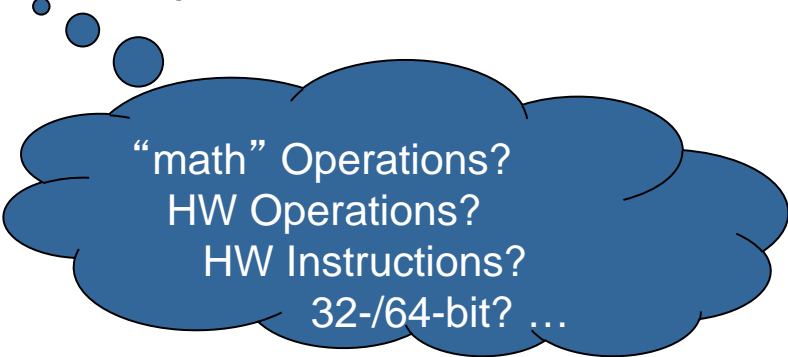
“If you optimize everything,  
you will always be unhappy.”

Donald E. Knuth



- What can be measured?
  - A **count** of how often an event occurs
    - E.g., the number of MPI point-to-point messages sent
  - The **duration** of some interval
    - E.g., the time spent these send calls
  - The **size** of some parameter
    - E.g., the number of bytes transmitted by these calls
- Derived metrics
  - E.g., rates / throughput
  - Needed for normalization

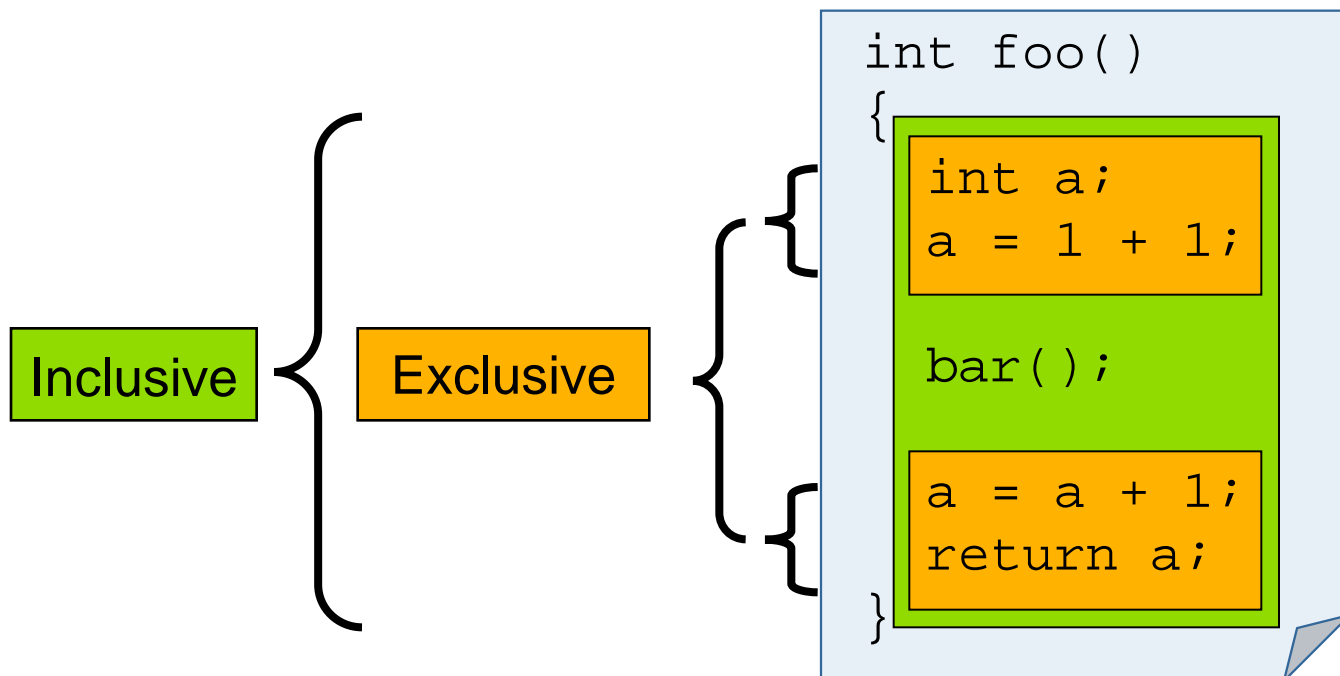
- Execution time
- Number of function calls
- CPI
  - CPU cycles per instruction
- FLOPS
  - Floating-point operations executed per second



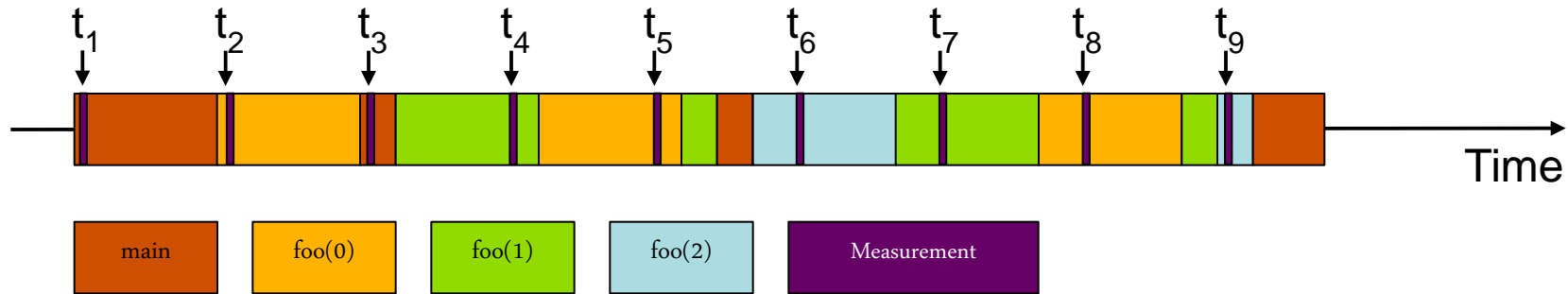
“math” Operations?  
HW Operations?  
HW Instructions?  
32-/64-bit? ...

- Wall-clock time
  - Includes waiting time: I/O, memory, other system activities
  - In time-sharing environments also the time consumed by other applications
- CPU time
  - Time spent by the CPU to execute the application
  - Does not include time the program was context-switched out
    - Problem: Does not include inherent waiting time (e.g., I/O)
    - Problem: Portability? What is user, what is system time?
- Problem: Execution time is non-deterministic
  - Use mean or minimum of several runs

- Inclusive
  - Information of all sub-elements aggregated into single value
- Exclusive
  - Information cannot be subdivided further



- How are performance measurements triggered?
  - Sampling
  - Code instrumentation
  
- How is performance data recorded?
  - Profiling / Runtime summarization
  - Tracing
  
- How is performance data analyzed?
  - Online
  - Post mortem



```

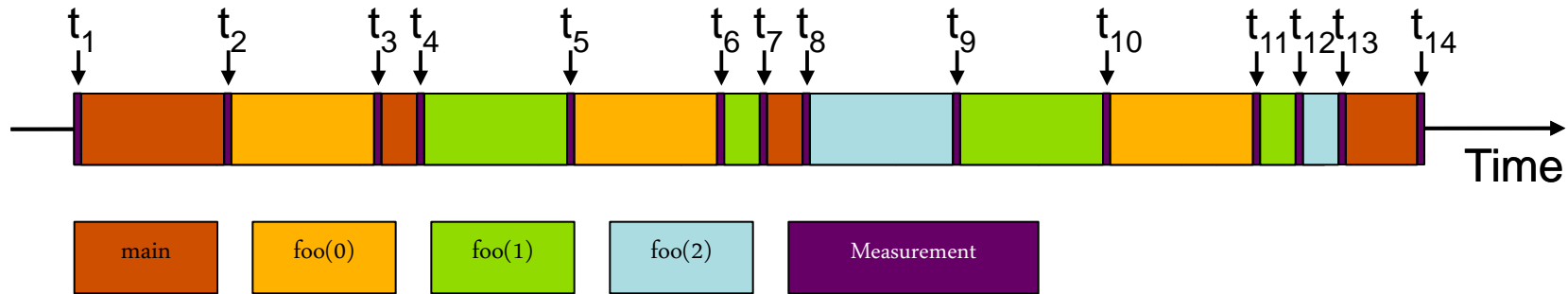
int main()
{
    int i;

    for (i=0; i < 3; i++)
        foo(i);

    return 0;
}

void foo(int i)
{
    if (i > 0)
        foo(i - 1);
}
    
```

- Running program is periodically interrupted to take measurement
  - Timer interrupt, OS signal, or HWC overflow
  - Service routine examines return-address stack
  - Addresses are mapped to routines using symbol table information
- **Statistical** inference of program behavior
  - Not very detailed information on highly volatile metrics
  - Requires long-running applications
- Works with unmodified executables



```

int main()
{
    int i;
    Enter("main");
    for (i=0; i < 3; i++)
        foo(i);
    Leave("main");
    return 0;
}

void foo(int i)
{
    Enter("foo");
    if (i > 0)
        foo(i - 1);
    Leave("foo");
}
    
```

- Measurement code is inserted such that every event of interest is captured **directly**
  - Can be done in various ways
- Advantage:
  - Much more detailed information
- Disadvantage:
  - Processing of source-code / executable necessary
  - Large relative overheads for small functions

- **Static** instrumentation
  - Program is instrumented prior to execution
- **Dynamic** instrumentation
  - Program is instrumented at runtime
- Code is inserted
  - Manually
  - Automatically
    - By a preprocessor / source-to-source translation tool
    - By a compiler
    - By linking against a pre-instrumented library / runtime system
    - By binary-rewrite / dynamic instrumentation tool



- Accuracy
  - Intrusion overhead
    - Measurement itself needs time and thus lowers performance
  - Perturbation
    - Measurement alters program behaviour
      - E.g., memory access pattern
  - Accuracy of timers & counters
- Granularity
  - How many measurements?
  - How much information / processing during each measurement?

☞ *Tradeoff: Accuracy vs. Expressiveness of data*

- How are performance measurements triggered?
  - Sampling
  - Code instrumentation
- How is performance data recorded?
  - Profiling / Runtime summarization
  - Tracing
- How is performance data analyzed?
  - Online
  - Post mortem

- Recording of aggregated information
  - Total, maximum, minimum, ...
- For measurements
  - Time
  - Counts
    - Function calls
    - Bytes transferred
    - Hardware counters
- Over program and system entities
  - Functions, call sites, basic blocks, loops, ...
  - Processes, threads

☞ *Profile = summarization of events over execution interval*

- Flat profile
  - Shows distribution of metrics per routine / instrumented region
  - Calling context is not taken into account
- Call-path profile
  - Shows distribution of metrics per executed call path
  - Sometimes only distinguished by partial calling context (e.g., two levels)
- Special-purpose profiles
  - Focus on specific aspects, e.g., MPI calls or OpenMP constructs
  - Comparing processes/threads

- Recording detailed information about significant points (events) during execution of the program
  - Enter / leave of a region (function, loop, ...)
  - Send / receive a message, ...
- Save information in event record
  - Timestamp, location, event type
  - Plus event-specific information (e.g., communicator, sender / receiver, ...)
- Abstract execution model on level of defined events

👉 *Event trace = Chronologically ordered sequence of event records*

# Event tracing

Process A

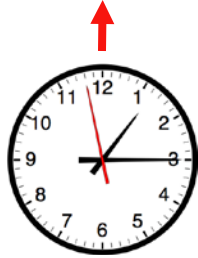
```
void foo() {  
  trc_enter("foo");  
  ...  
  trc_send(B);  
  send(B, tag, buf);  
  ...  
  trc_exit("foo");  
}
```

instrument

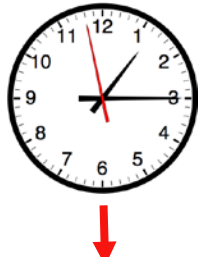
Process B

```
void bar() {  
  trc_enter("bar");  
  ...  
  recv(A, tag, buf);  
  trc_recv(A);  
  ...  
  trc_exit("bar");  
}
```

MONITOR



synchronize(d)



MONITOR

Local trace A

...	
58	ENTER foo
62	SEND to B
64	EXIT foo
...	

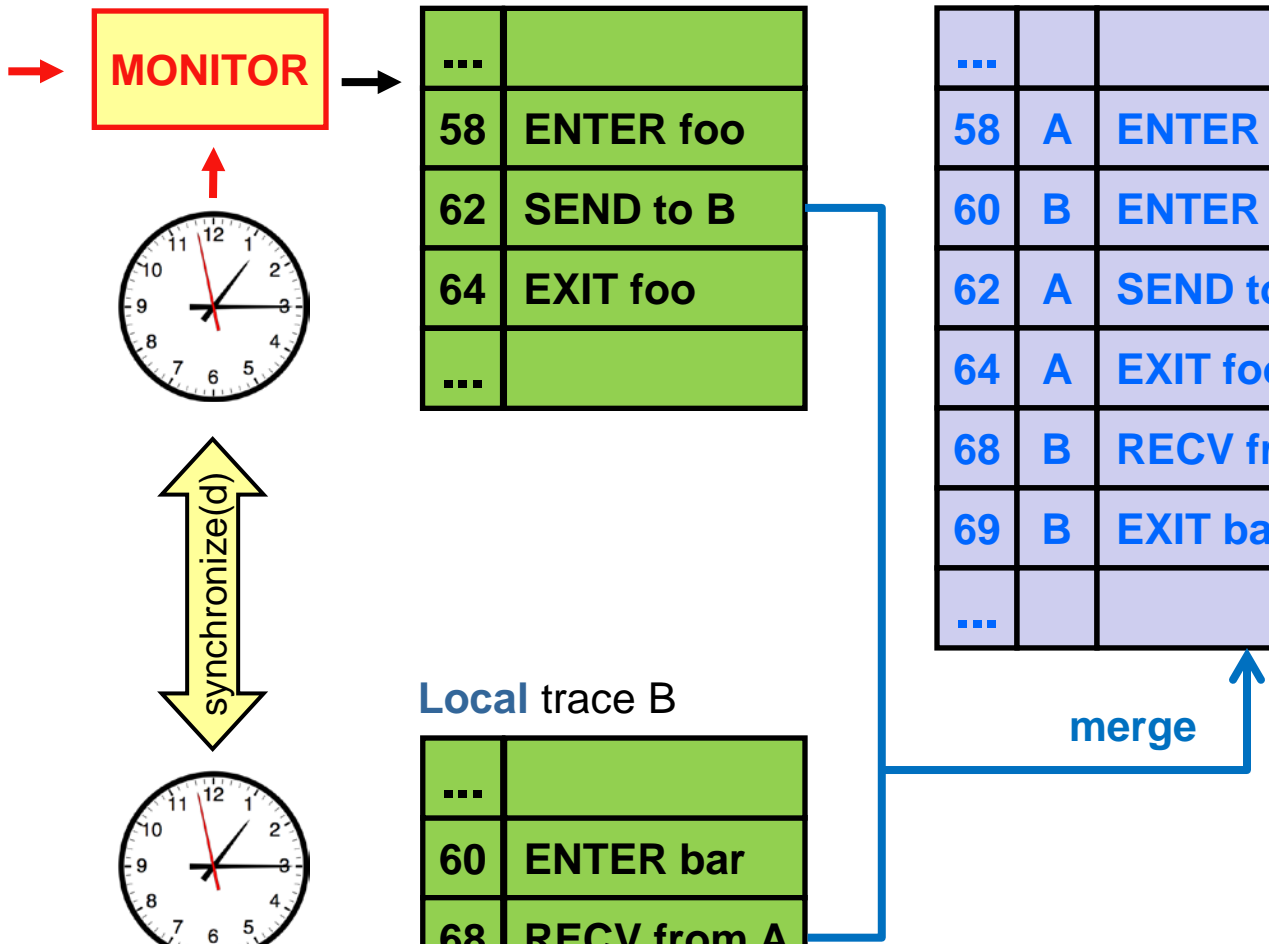
Local trace B

...	
60	ENTER bar
68	RECV from A
69	EXIT bar
...	

Global trace view

...		
58	A	ENTER foo
60	B	ENTER bar
62	A	SEND to B
64	A	EXIT foo
68	B	RECV from A
69	B	EXIT bar
...		

merge



## ■ Tracing advantages

- Event traces preserve the **temporal** and **spatial** relationships among individual events (👉 context)
- Allows reconstruction of **dynamic** application behaviour on any required level of abstraction
- Most general measurement technique
  - Profile data can be reconstructed from event traces

## ■ Disadvantages

- Traces can very quickly become extremely large
- Writing events to file at runtime may causes perturbation

- How are performance measurements triggered?
  - Sampling
  - Code instrumentation
- How is performance data recorded?
  - Profiling / Runtime summarization
  - Tracing
- How is performance data analyzed?
  - Online
  - Post mortem



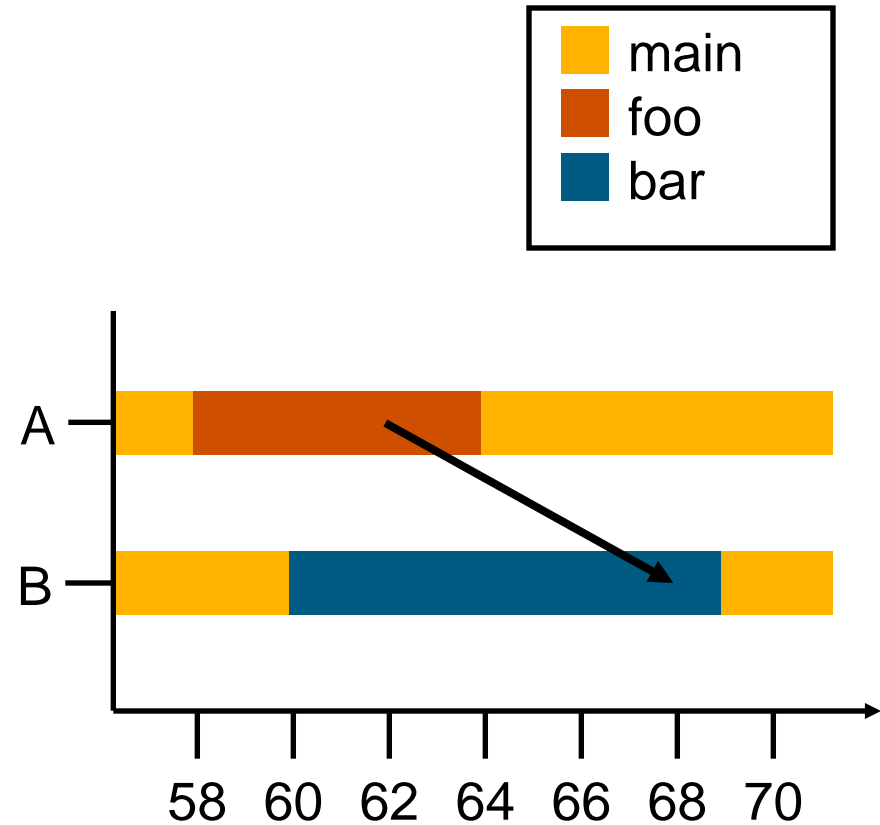
- Performance data is processed during measurement run
  - Process-local profile aggregation
  - More sophisticated inter-process analysis using
    - “Piggyback” messages
    - Hierarchical network of analysis agents
- Inter-process analysis often involves application steering to interrupt and re-configure the measurement

- Performance data is stored at end of measurement run
- Data analysis is performed afterwards
  - Automatic search for bottlenecks
  - Visual trace analysis
  - Calculation of statistics

## Global trace view

...		
58	A	ENTER foo
60	B	ENTER bar
62	A	SEND to B
64	A	EXIT foo
68	B	RCV from A
69	B	EXIT bar
...		

Post-Mortem  
 Analysis





☞ *A combination of different methods, tools and techniques is typically needed!*

- Analysis
  - Statistics, visualization, automatic analysis, data mining, ...
- Measurement
  - Sampling / instrumentation, profiling / tracing, ...
- Instrumentation
  - Source code / binary, manual / automatic, ...

- Do I have a performance problem at all?
  - Time / speedup / scalability measurements
- **What** is the key bottleneck (computation / communication)?
  - MPI / OpenMP / flat profiling
- **Where** is the key bottleneck?
  - Call-path profiling, detailed basic block profiling
- **Why** is it there?
  - Hardware counter analysis, trace selected parts to keep trace size manageable
- Does the code have scalability problems?
  - Load imbalance analysis, compare profiles at various sizes function-by-function

- What is executed, how often, how long?
  - what execution sequence, what context?
  - what is executed serially vs. parallel, what proportion of time?
    - which threads, which cores?
- What are the OpenMP parallelisation overheads?
  - thread management (fork/join of parallel regions)
  - synchronisations (implicit and explicit)
    - barriers, critical/atomic, ordered, locks (taskwait)
  - memory flushes, task creations
- What are inefficient parallelisations?
  - imbalance, limited parallelism
- What other factors are involved?
  - resource contention, oversubscription, binding
  - data placement & locality