



# POP3: Services and Methodology

Marta Garcia-Gasulla, BSC

HORIZON-EUROHPC-JU-2023-COE



**EuroHPC**  
Joint Undertaking

1 January 2024– 31 December 2026

Grant Agreement No 101143931

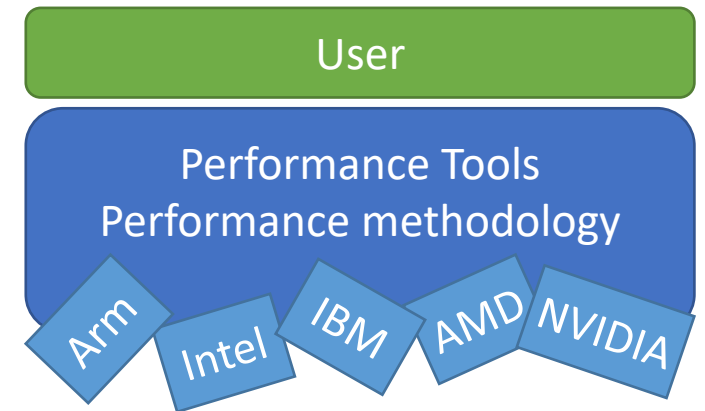


- A **Centre of Excellence**
  - On **Performance Optimisation and Productivity**
  - Promoting **best practices in parallel programming**
- Providing **FREE Services**
  - Precise understanding of application and system behaviour
  - Suggestion/support on how to refactor code in the most productive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

# Context and history



- **POP-CoE** starts in 2015
- Objectives:
  - To offer performance analysis services to European HPC users.
  - To develop a common methodology for performance analysis
- **POP1** runs from October 2015 to March 2018
- **POP2** runs from December 2018 to June 2022
- **POP3** will run between January 2023 and December 2026



# Partners



## • Who?

- BSC, ES (coordinator)
- HLRS, DE
- INESC-ID, PT
- IT4I, CZ
- JSC, DE
- 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

# Partners



## • Who?

- BSC, ES (coordinator)
- HLRS, DE
- INESC-ID, PT
- IT4I, CZ
- JSC, DE
- RWTH Aachen, IT Center, DE
- TERATEC, FR
- UVSQ, FR



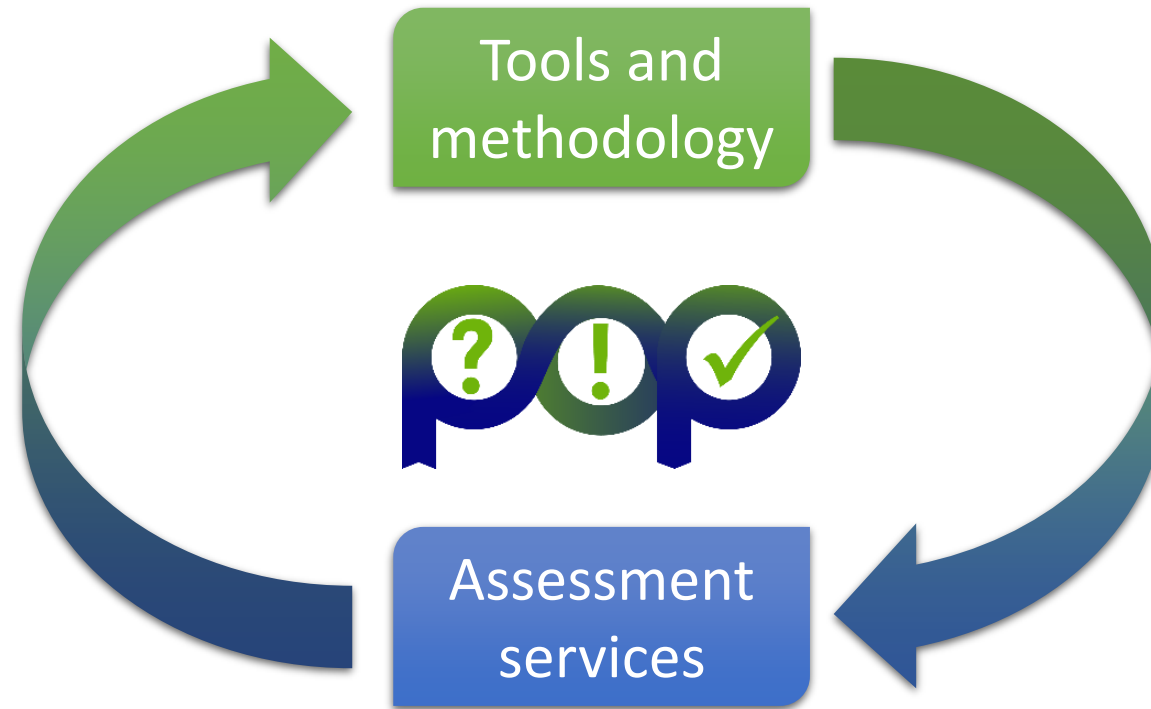
## Why it works?

- Diversity of expertise to cover all areas
  - Some of them tool developers, some of them analysts, some customer advocate
- Diversity of tools for resiliency
- Diversity of backgrounds to design a common methodology

# Objectives



# Objectives





# POP3 Services



- **Parallel Application Performance Assessment**

- Primary service
- Initial analysis measuring a [range of performance metrics](#) to assess quality of performance and identify the issues affecting performance (at customer site)
- If needed, undertakes further performance evaluations to identify the root causes of the issues found and qualify and quantify approaches to address them (recommendations)

- **Second Level Services**

- Second level services may follow after conclusion of an initial performance assessment:
  - **Proof-of-concept**: explore the potential benefit of proposed optimisations by applying them to selected regions of the applications
  - **Correctness-check**: evaluate the correctness of hybrid MPI + OpenMP applications
  - **Energy-efficiency study**: investigate improvements of energy consumption or efficiency
  - **Advisory study**: ongoing consultancy for customers that choose to implement proposed optimisations on their own

- **Note: Effort shared between our experts and customer!**

# 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

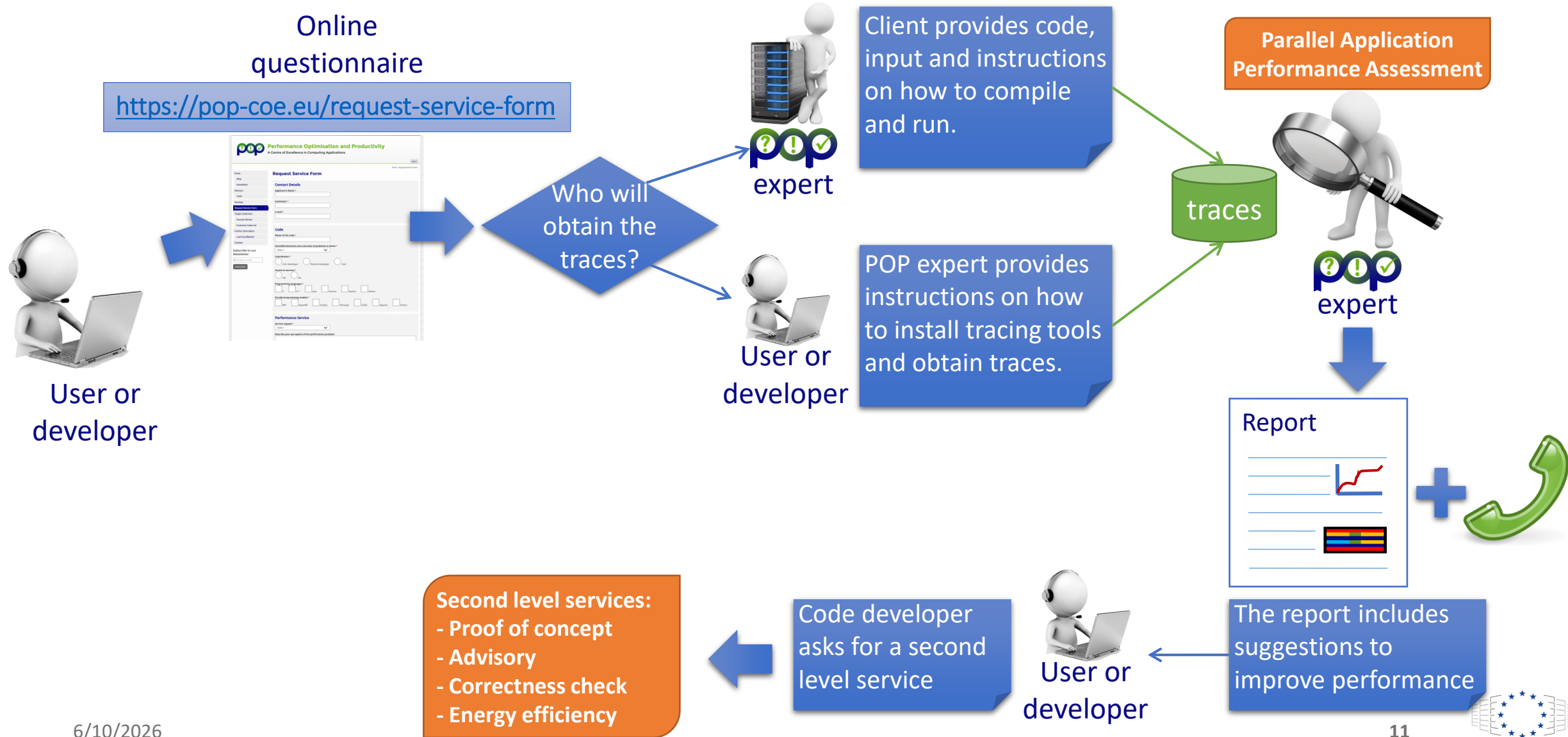
- **Infrastructure operators**

- 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

# POP3 services



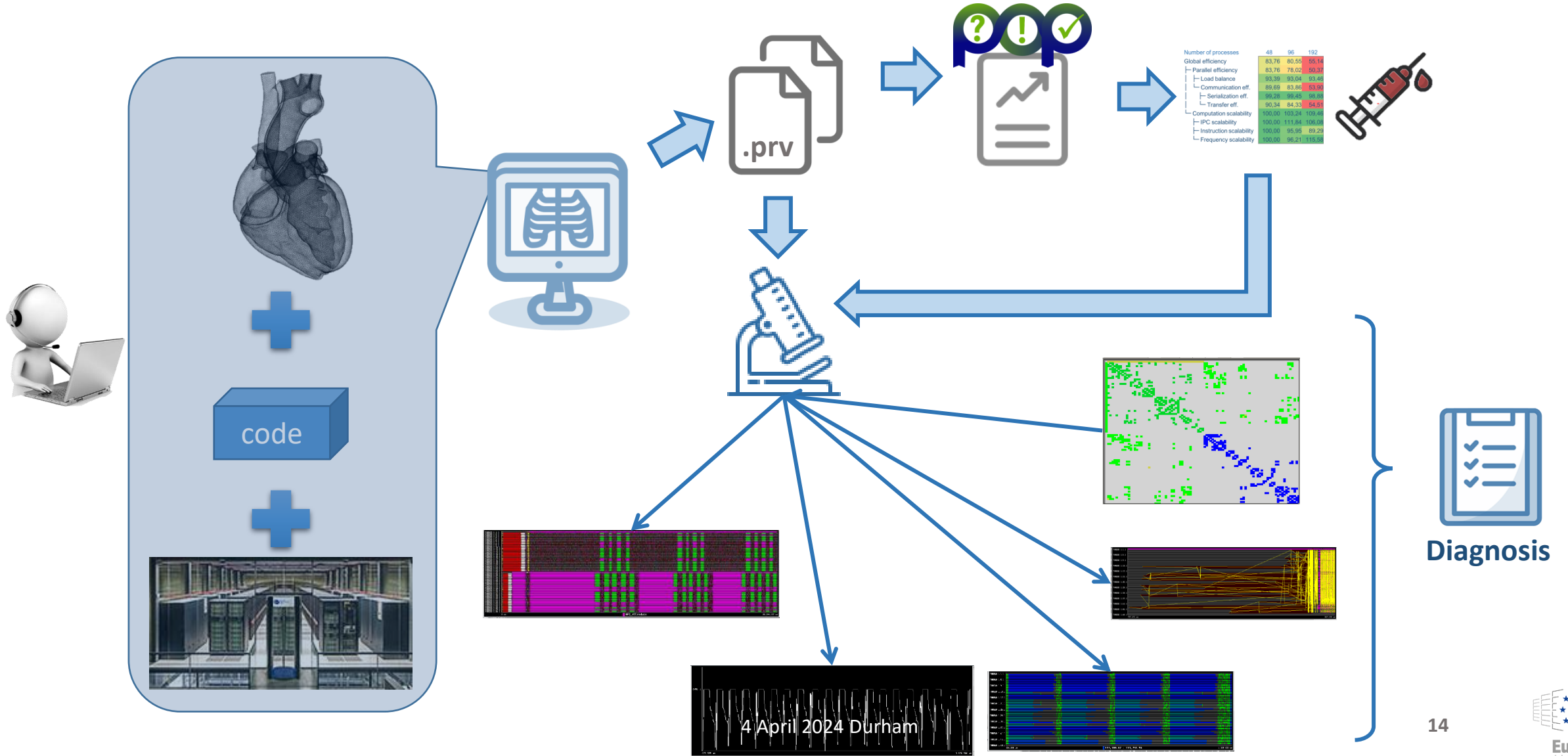


# POP3 Methodology



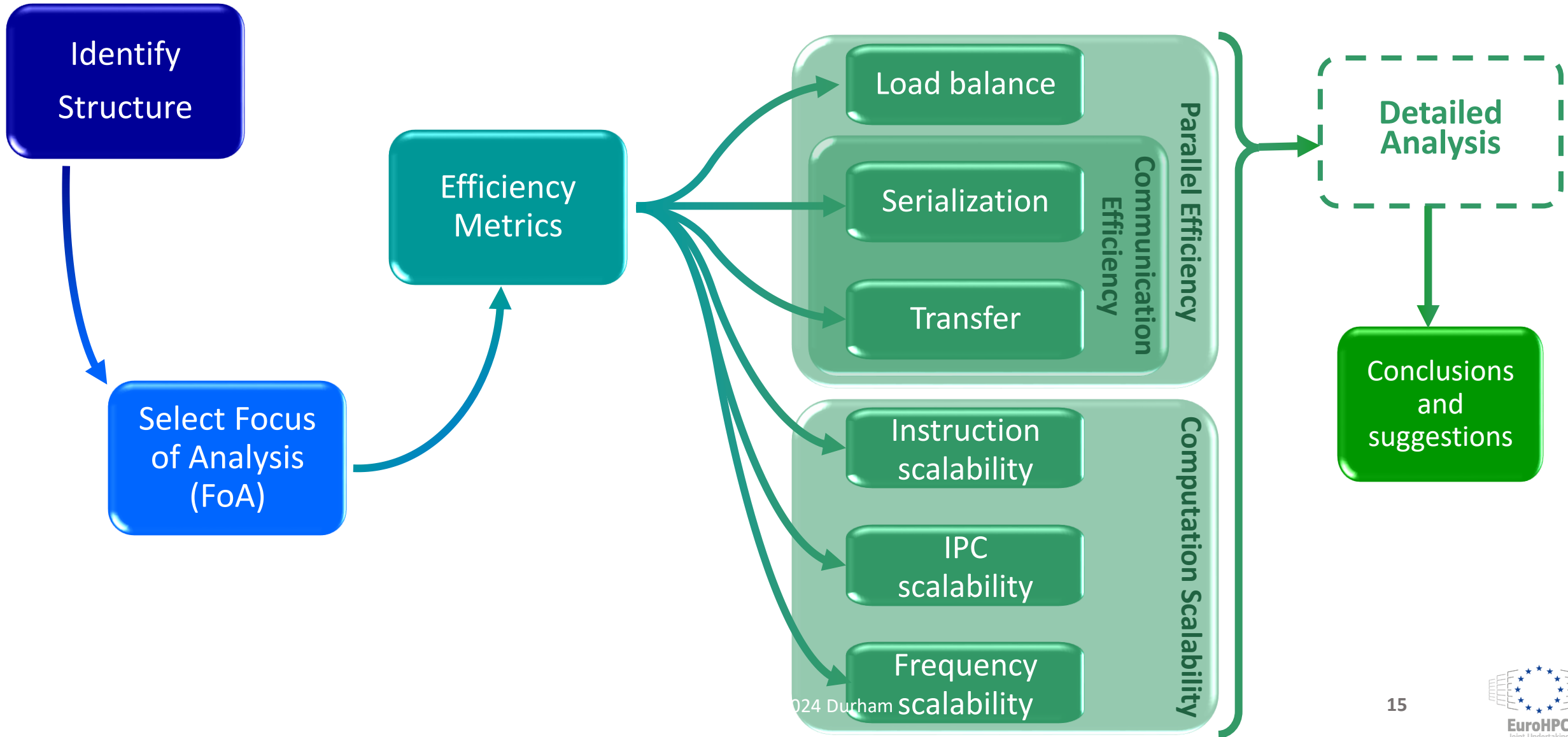
- POP methodology
  - A methodology for performance analysis
  - Agnostic of the tools
    - But this presentation is based on BSC tools (Extrae, Paraver, Dimemas and ModelFactors)
    - You will see examples of the methodology applied using JSC tools on 25th April 2024
- MPI only applications
  - But POP methodology is being extended to:
    - Hybrid models: OpenMP, GPUs...
    - I/O
    - Vectorization

# Performance analysis

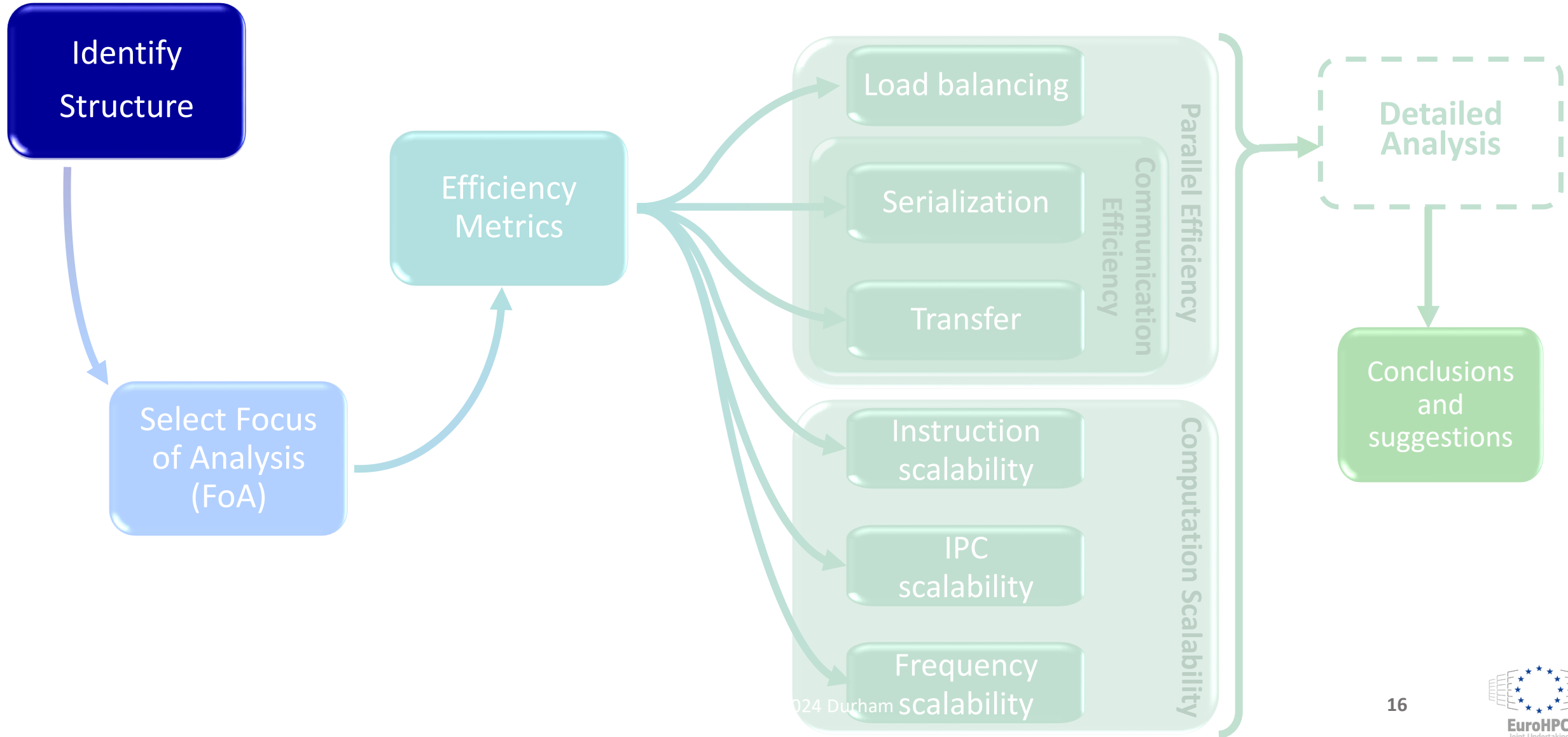


4 April 2024 Durham

# The journey



# The journey



# Identify structure

---

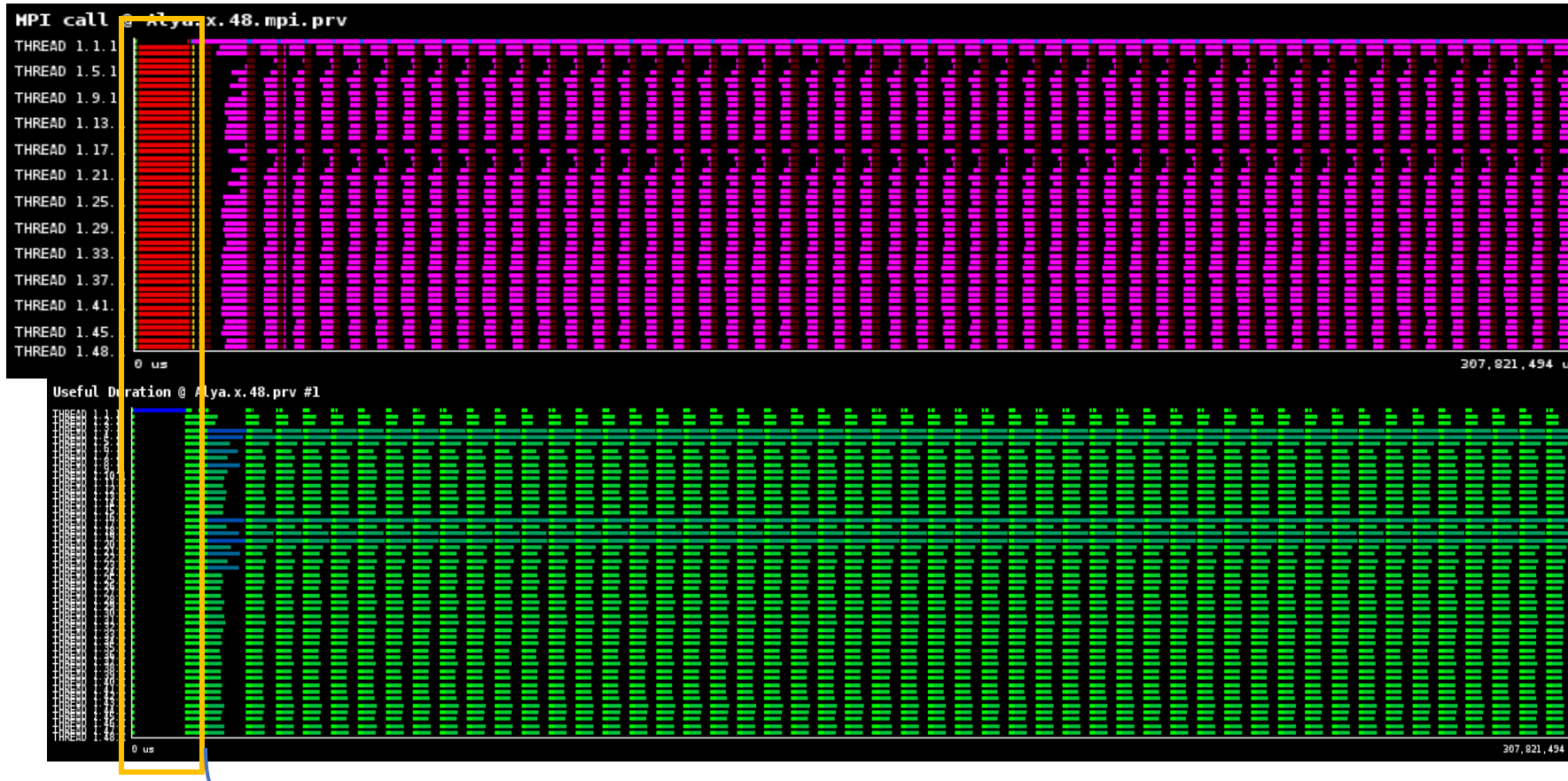


- Objectives:
  - Understand general structure
  - Identify initialization/finalization phases
  - Detect iterative pattern
  - Granularity
- Different levels of “difficulty” → Different levels of knowledge

# Identify structure



- Usually use “MPI calls” view or “Useful Duration”



Initialization

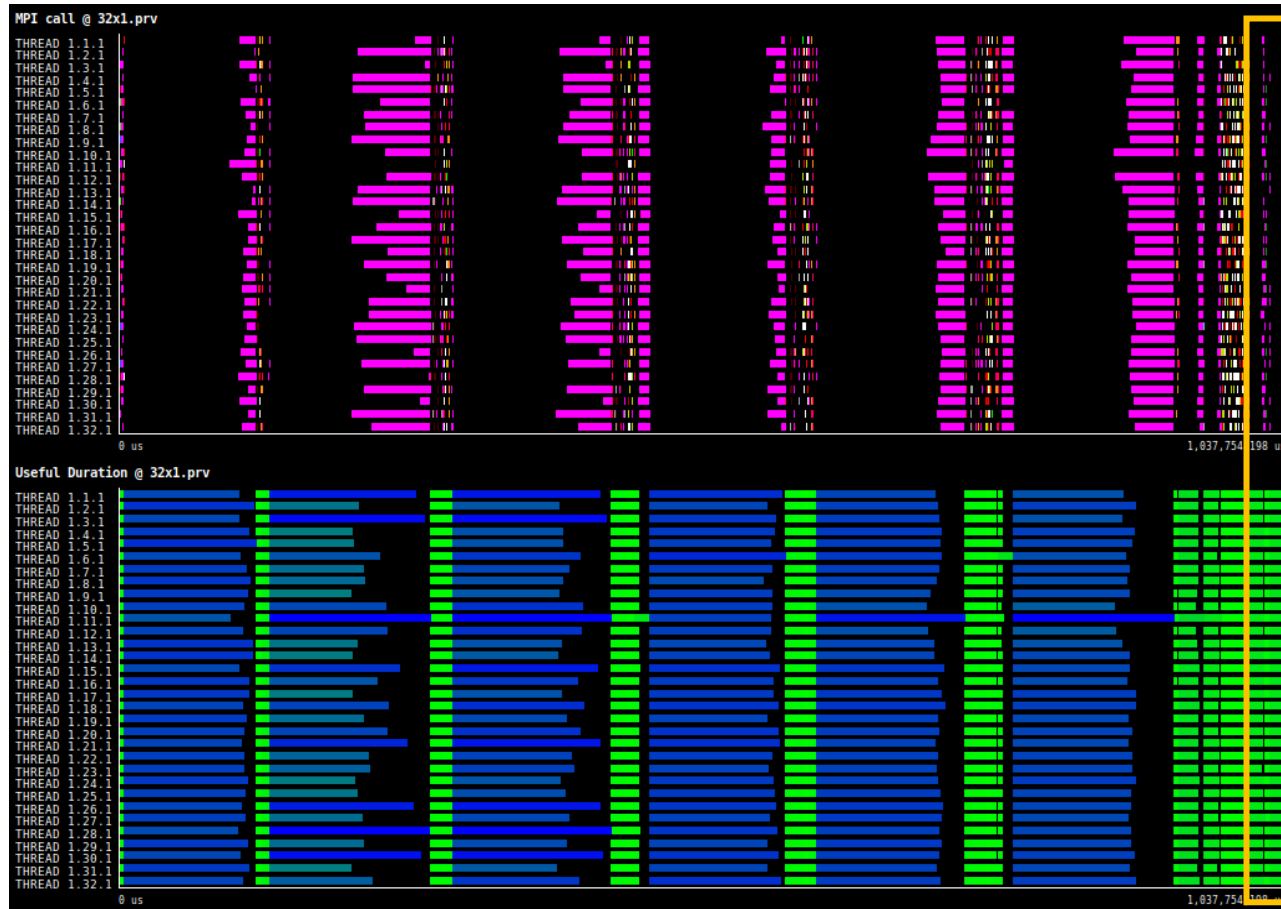
Iterative phase

- Clear iterative pattern
- With an initialization phase
- All iterations are similar
  - We can select a few to analyze

# Identify structure



- Not always easy



6 iterations

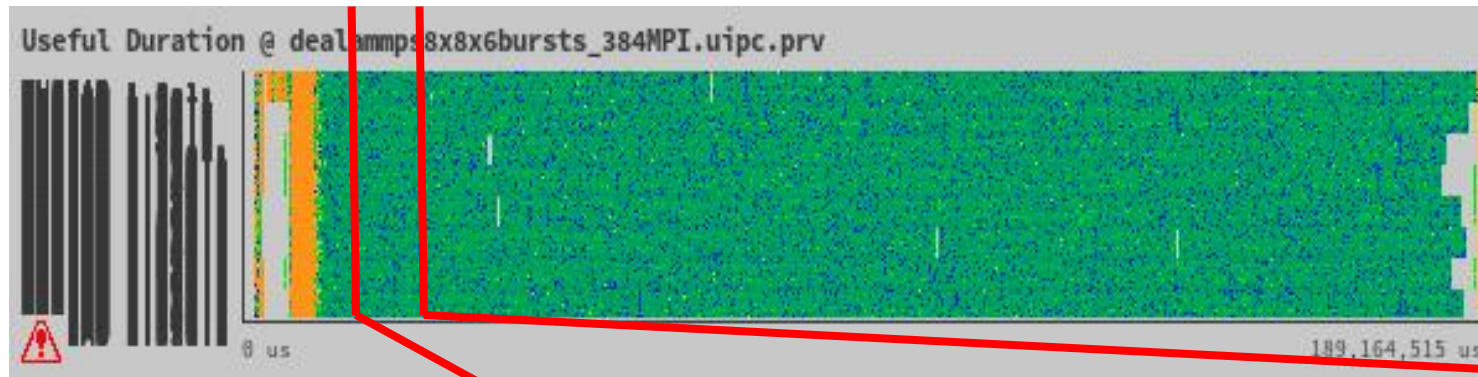
Finalization

- There are 6 iterations and one finalization phase
- Iterations are not regular along time
  - Different pattern of load balance
  - Different pattern of durations

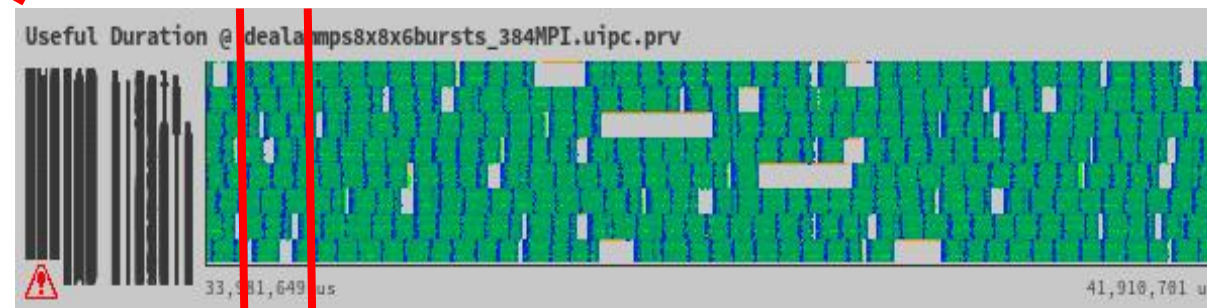
# Identify structure



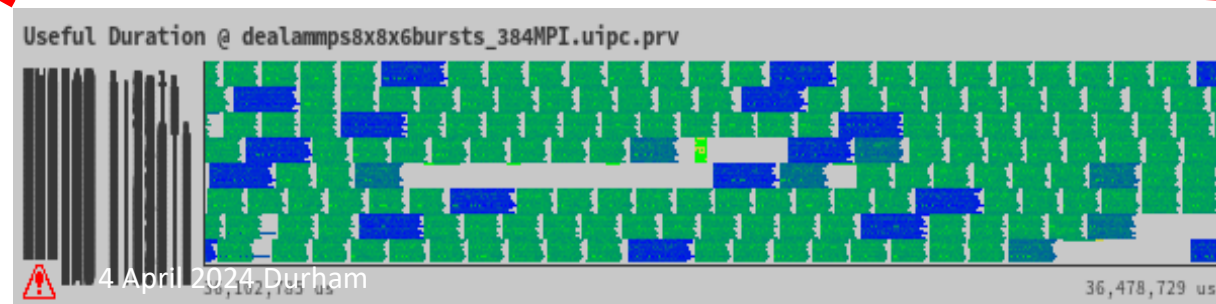
- Not always easy



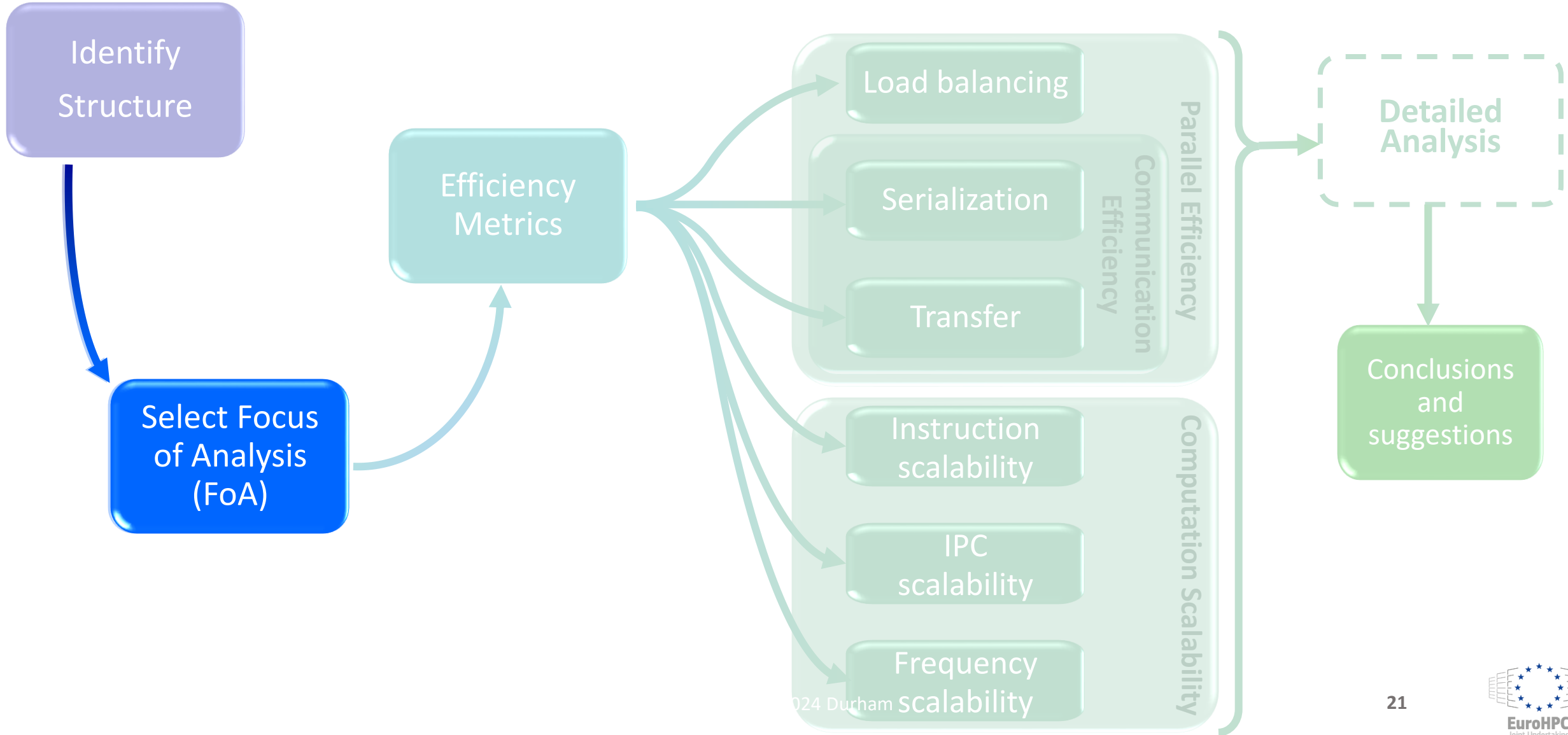
- It is not easy to detect an iterative structure
- No global synchronizations



- Zooming in we detect that synchronizations only happen at node level
- Not possible to determine iterations



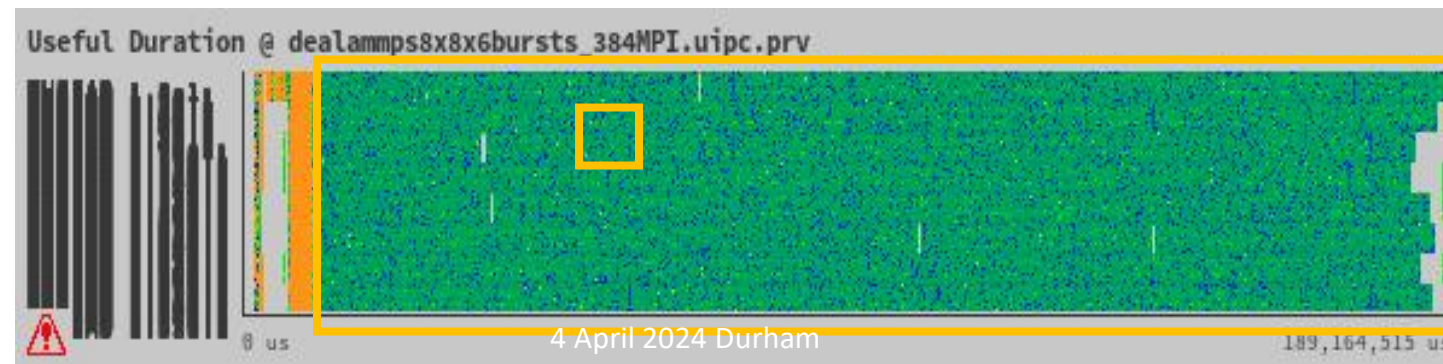
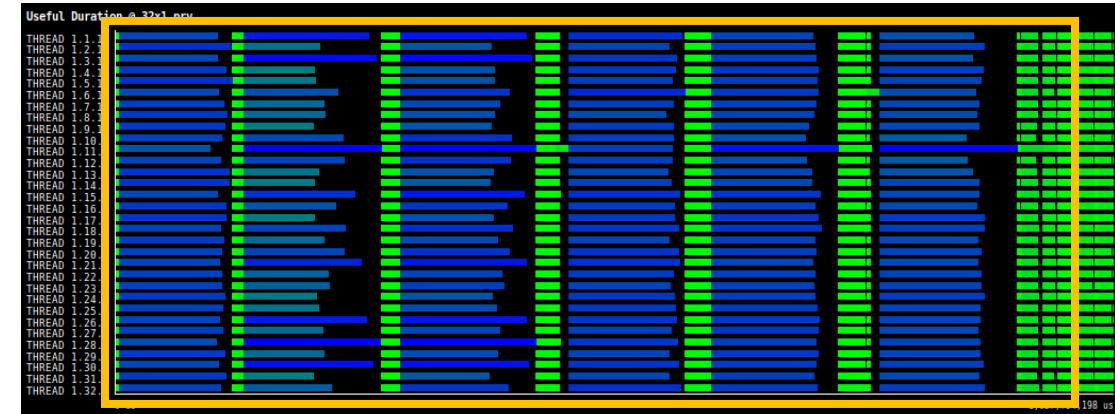
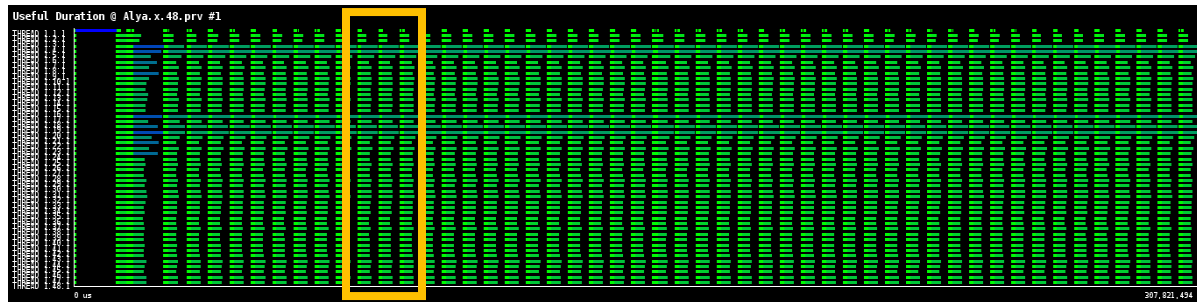
# The journey



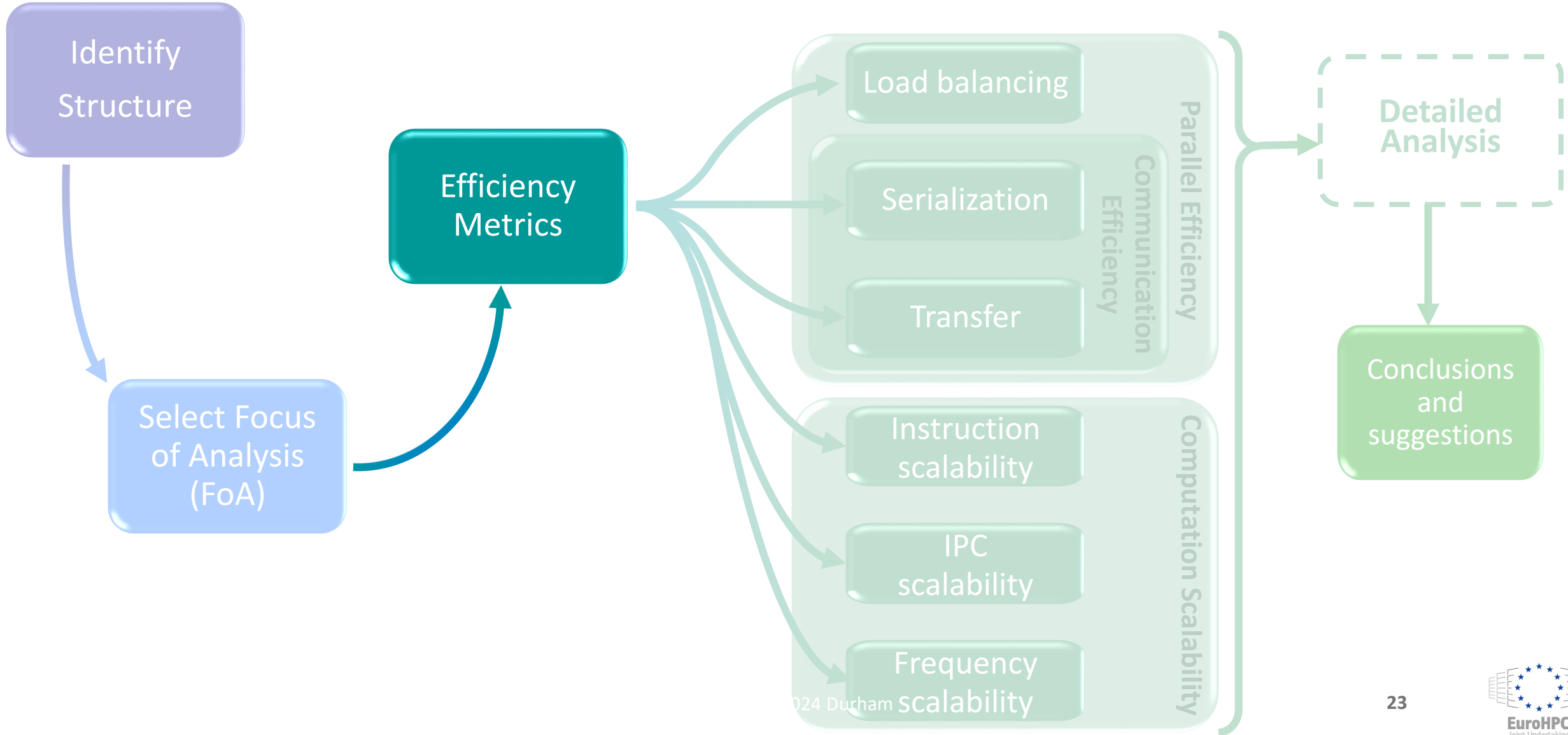
# Select Focus of Analysis (FoA)



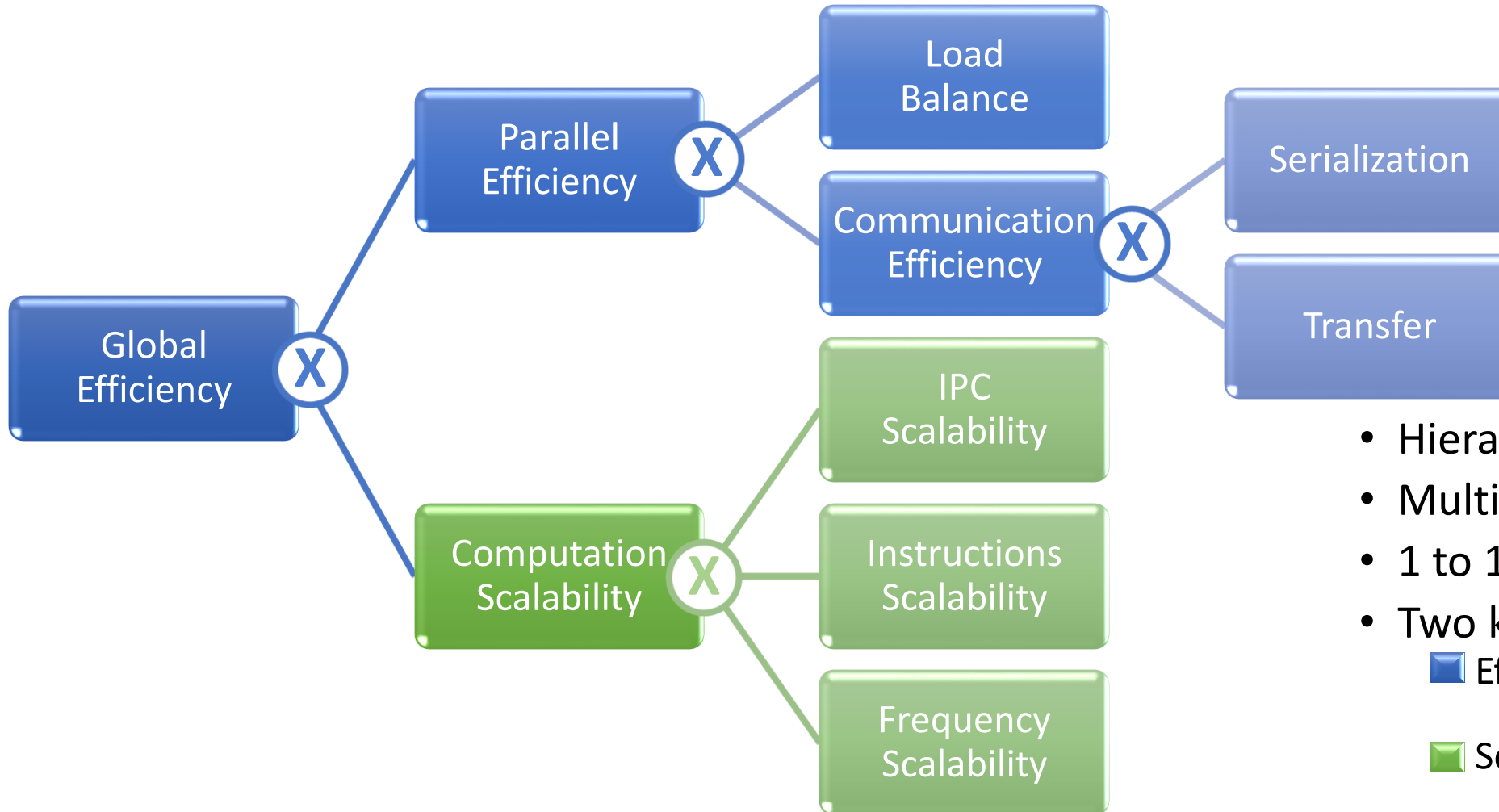
- **Objective:** Select the region we want to analyze
  - Not a correct answer, depends on the context of the analysis
  - For the same trace we may select two different FoA to perform two different studies with two different objectives.



# The journey



# The Efficiency Metrics



- Hierarchical model
- Multiplicative metrics
- 1 to 100% scale
- Two kind of metrics:
  - Efficiency metrics
    - Absolute metrics
  - Scalability metrics
    - Relative to a base case
    - 100% for the base case

# The Efficiency Metrics

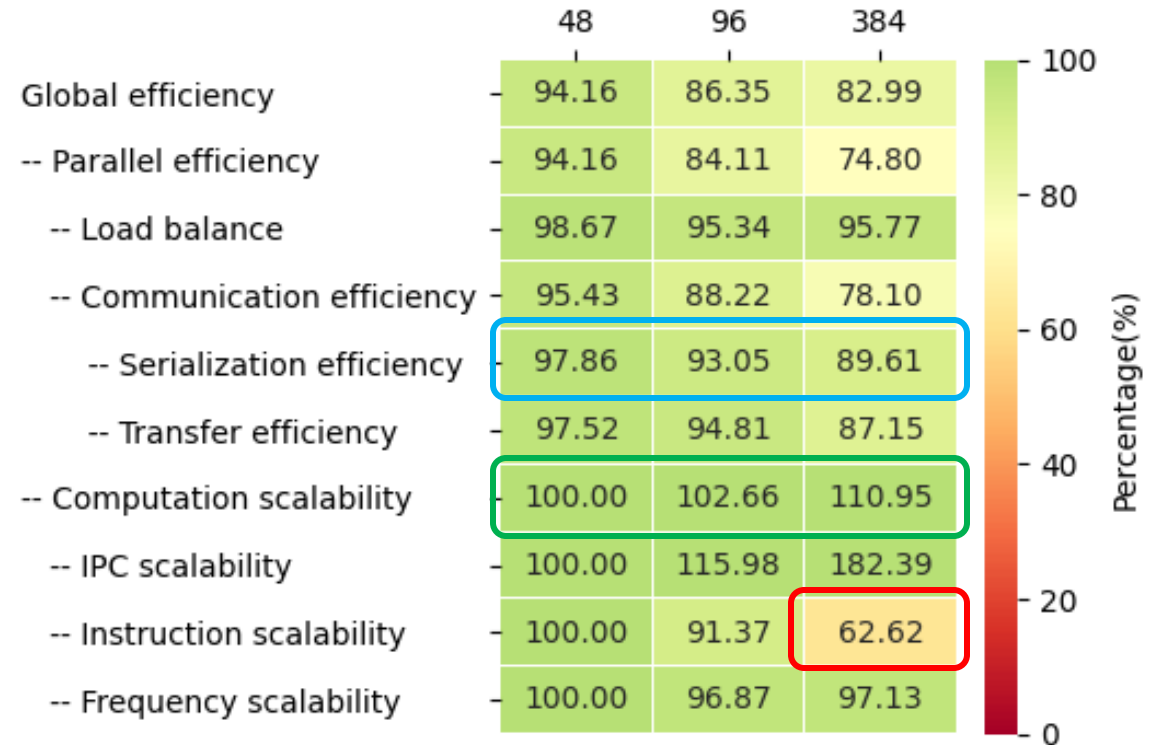


- What they are NOT
  - The end of the journey
- What they are...
  - ... a general mechanism to describe the fundamental concepts of parallelism
  - ... a hint that tells where to look
  - ... a way to quantify efficiency loss
  - ... a fair comparison between different...
    - ... code versions
    - ... architectures
    - ... core counts (scalability)
    - ... applications
- For each metric we are going to see:
  - What it quantifies
  - How it is computed
    - Formula
    - Graphical
  - Interpretation
  - Where to look next

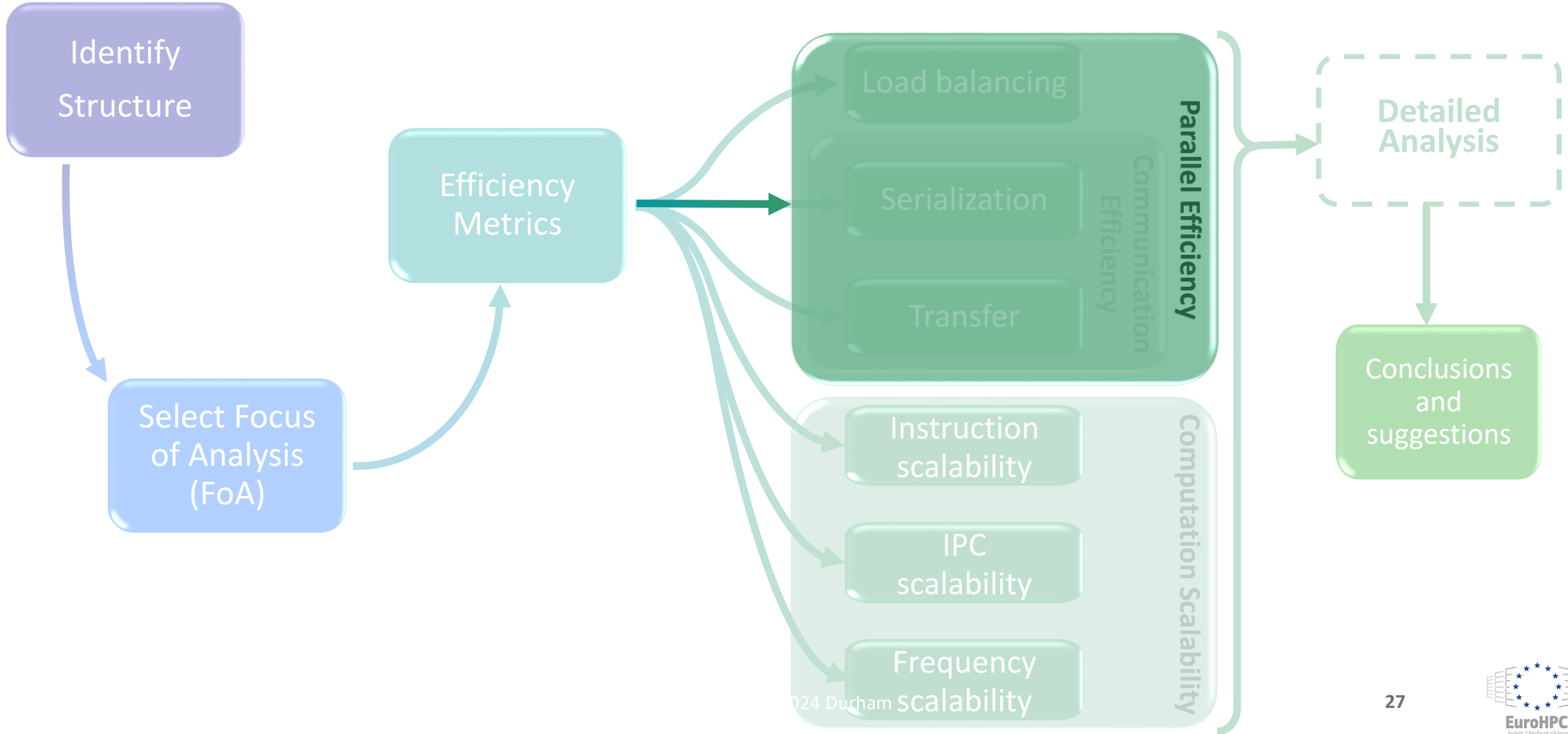
# The Efficiency Metrics



- Usually shown in a table
  - Rows: Metrics
  - Columns: Different traces
- With colored cells as a heat map
- What to look for?
  - Low values
  - Trend
  - High values



# The journey

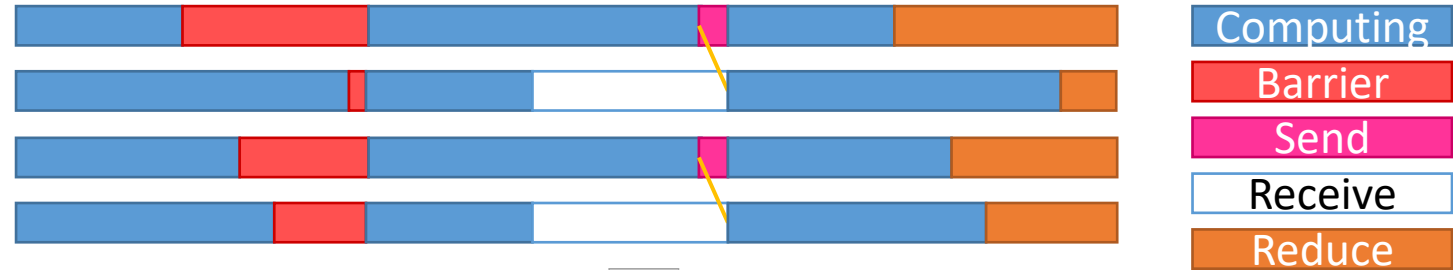


# Some semantics first



- The state of a processes is simplified to two values:

- Useful == Computing
- Not useful == Otherwise

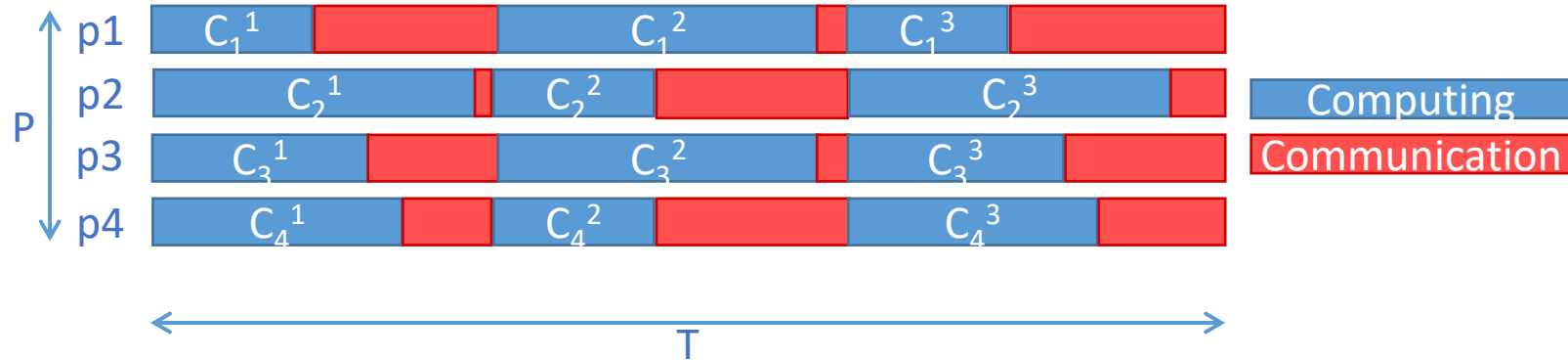


- T = Elapsed time
- P = Number of processes
- $c_i$  = Compute time of process i

- $c_i = \sum_{j=1}^n c_i^j$

- C = Total compute time

- $C = \sum_{i=1}^P c_i$



# Parallel Efficiency



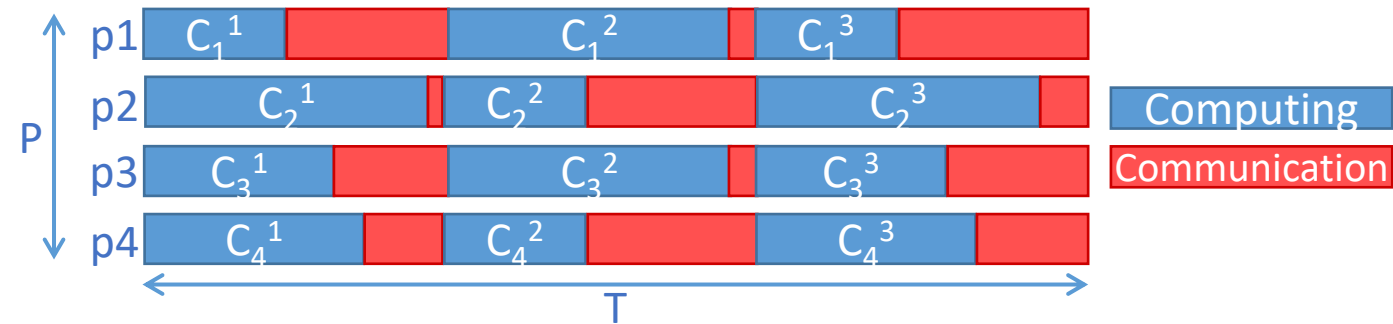
**Quantifies:** The extent to which all resources in the system are kept active doing useful work

**How it is computed:** Ratio between time used to do useful computation and consumed cpu time

$$\text{Parallel Efficiency} = \frac{\text{■}}{(\text{■} + \text{■})}$$

**How it is computed:**

$$\text{Parallel Efficiency} = \frac{\sum_{i=1}^P c_i}{P * T}$$

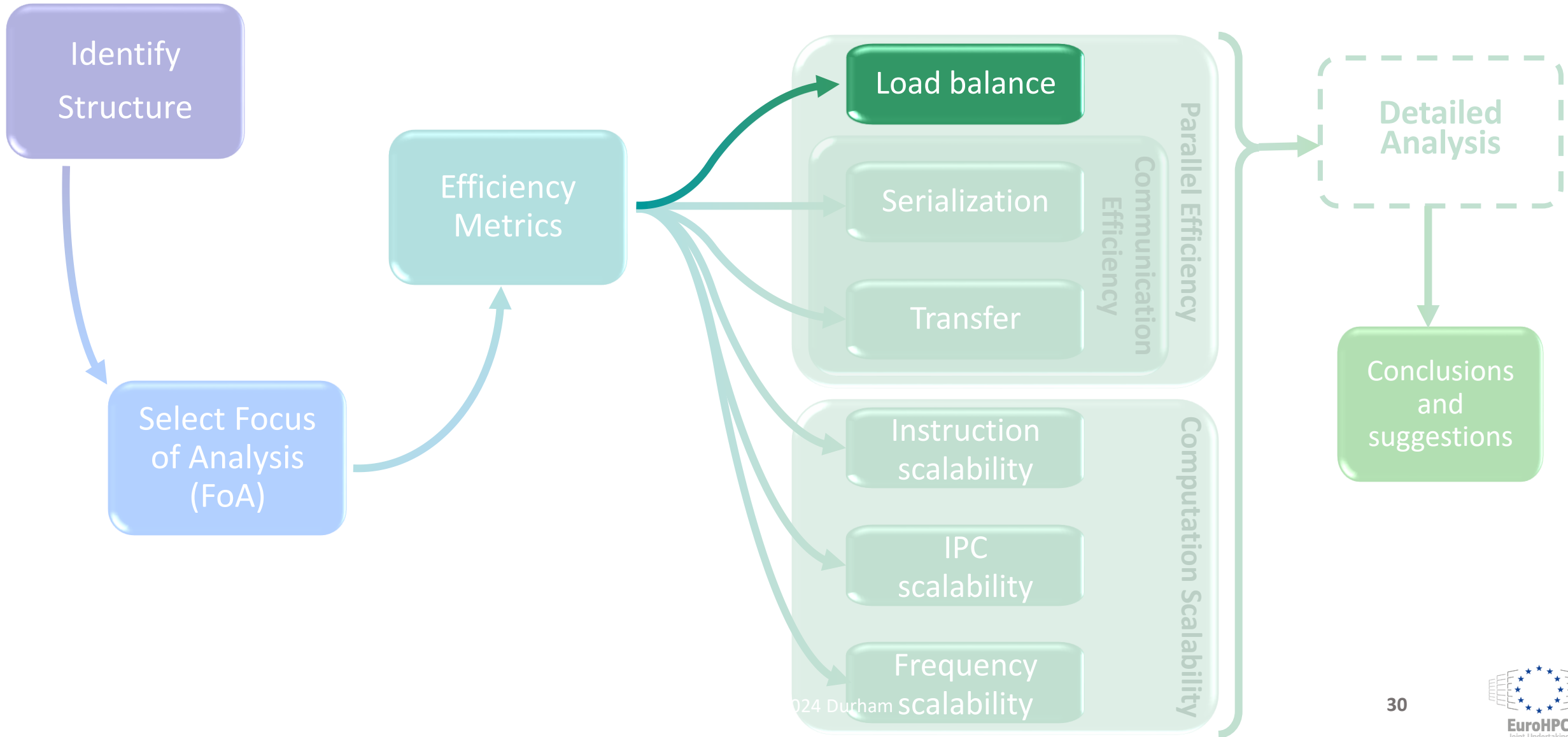


**Interpretation:** A low value indicates that a low fraction of the time consumed is used to do useful computation.

**Where to look next?**

- Look at its child metrics

# The journey



# Load Balance



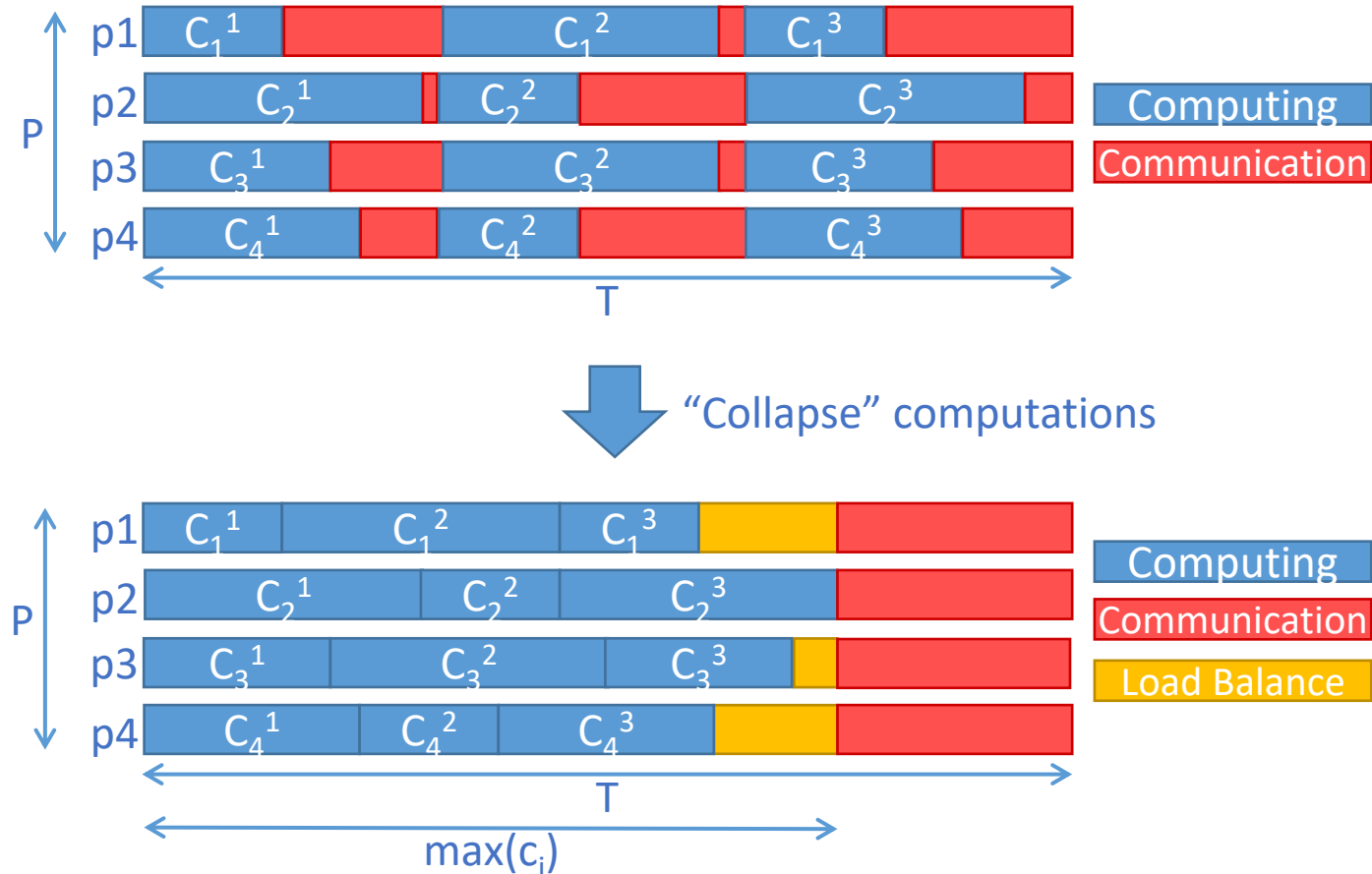
**Quantifies:** The efficiency loss due to the global distribution of work among processes.

**How it is computed:** Ratio between time used to do useful computation and the useful computation of the most loaded process multiplied by the number of processes

$$\text{Load Balance} = \frac{\text{blue}}{\text{blue} + \text{yellow}}$$

**How it is computed:**

$$\text{Load Balance} = \frac{\sum_{i=1}^P C_i}{\max_{i=1}^P (c_i) * P}$$

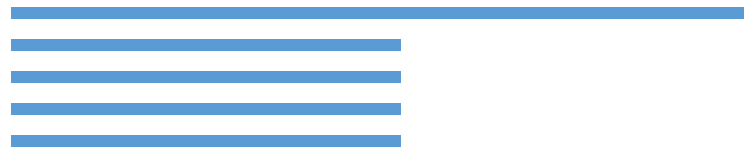


# Load Balance



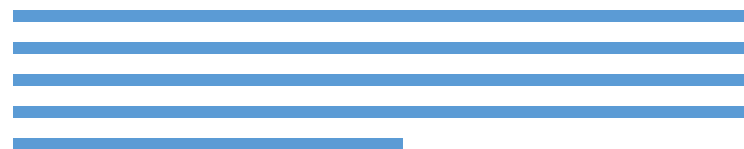
**Interpretation:** A low value in this metric indicates that more highly loaded processes keep other processes idle for a significant amount of time.

- A single highly loaded process will make this metric report a low value



$$LB = \frac{2 + 1 + 1 + 1 + 1}{2 * 5} = \frac{6}{10} = 0.6$$

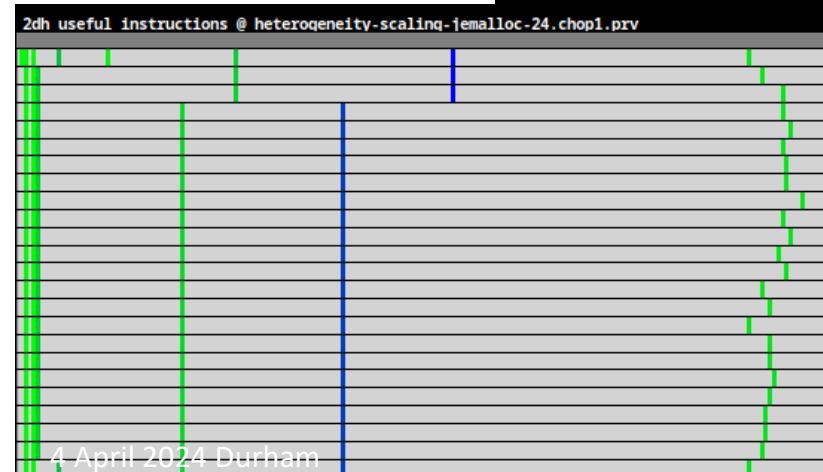
- A single low loaded process won't have an effect on this metric



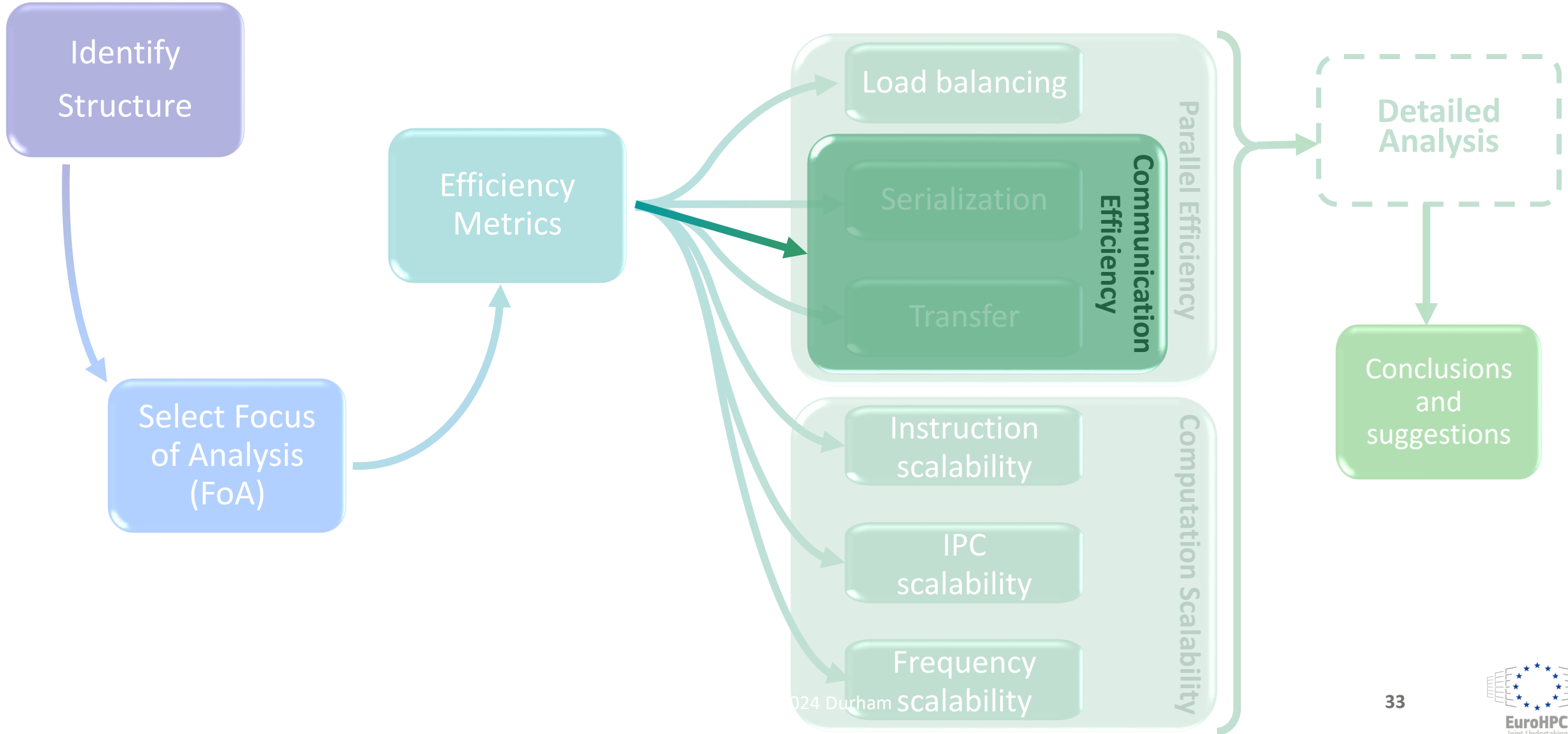
$$LB = \frac{2 + 2 + 2 + 2 + 1}{2 * 5} = \frac{9}{10} = 0.9$$

## • What to look next?

- We try to understand the cause of the Load Imbalance
- In general can have 3 sources
  - Number of instructions → Histogram of useful instructions
  - IPC → Histogram of IPC
  - Frequency → Histogram of cycles per us



# The journey



# Communication Efficiency



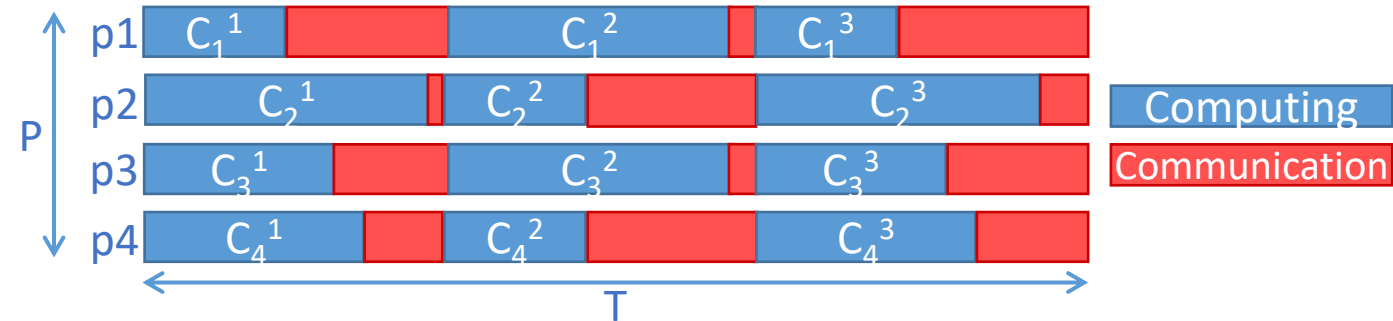
**Quantifies:** The efficiency loss due to the communication of data. Be it due to synchronizations between processes or to the overhead introduced by the communication itself. Excluding time loss due to global load imbalance

**How it is computed:** Ratio between the useful computation time of the most loaded processes and the total elapsed time

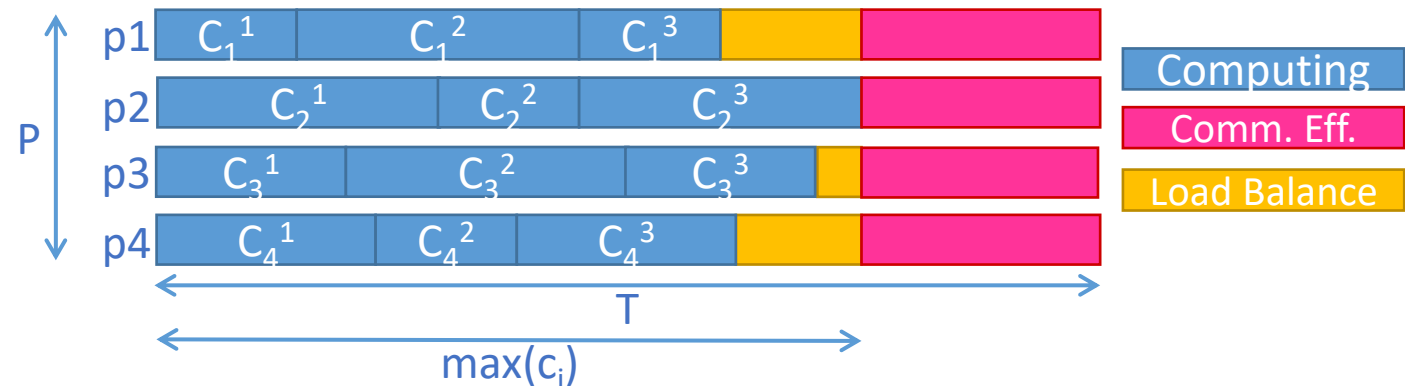
Communication Eff. = (  +  ) / (  +  +  )

**How it is computed:**

$$\text{Communication Eff.} = \frac{\max_{i=1}^P(c_i)}{T}$$



“Collapse” computations

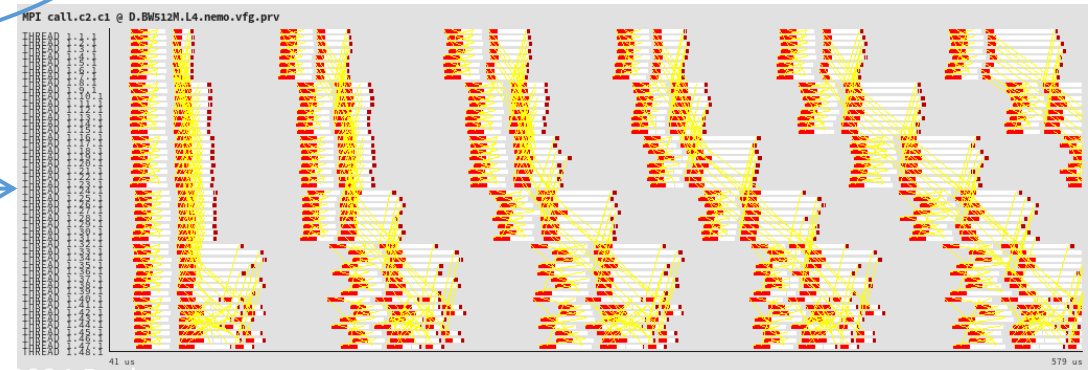
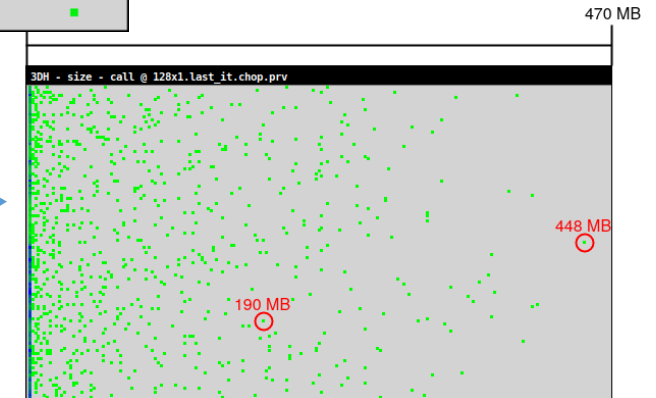


# Communication Efficiency

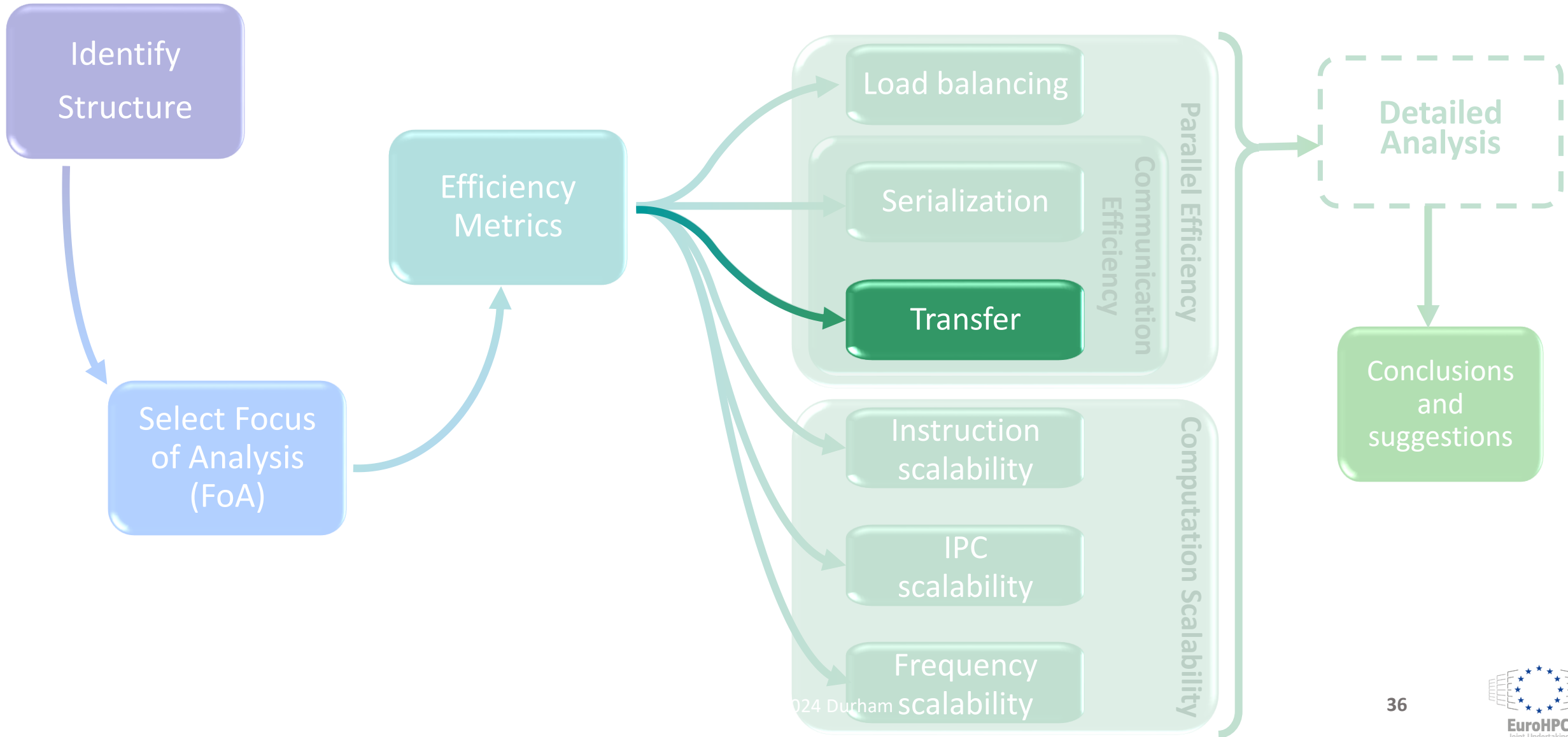


**Interpretation:** A low value in this metric indicates that the interaction between processes is impacting the performance.

- This metric can report good values in codes where a profile reports significant time in MPI and that would be reported by a bad load balance efficiency
- What to look next?
- If possible child metrics
- If not...
  - How many MPI calls are made?
    - Histogram of MPI calls
  - How often? Granularity of computations
    - Useful duration
  - How much data is sent?
    - Bytes sent per MPI call
  - Which are the semantics of the communication?
    - Chains of dependences?
      - MPI calls



# The journey



# Transfer Efficiency



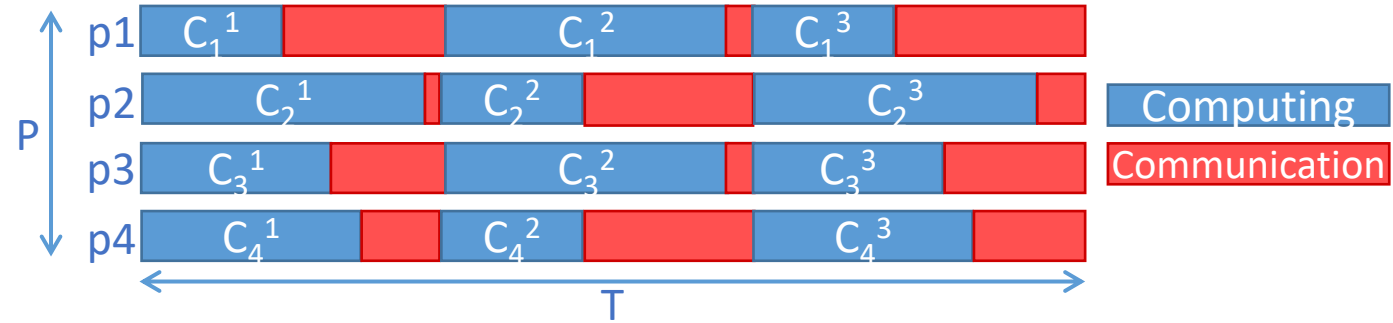
**Quantifies:** Efficiency loss related to the non instantaneous nature of communication mechanisms. Includes time to transmit the data over the physical channel and the overhead in the libraries.

**How it is computed:** Ratio between elapsed time in the ideal simulation and the elapsed time in the real execution

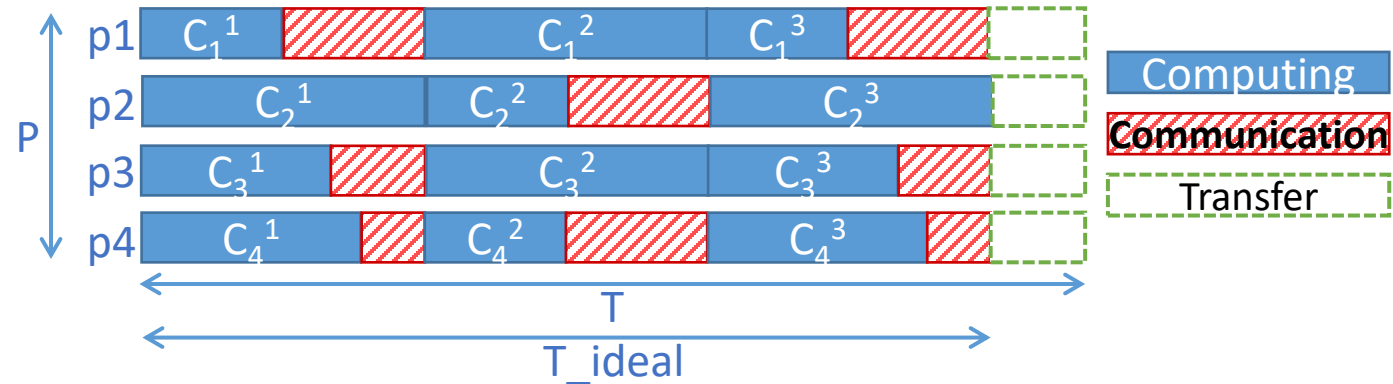
$$\text{Transfer Eff.} = \frac{\text{[Blue]} + \text{[Red Hatched]}}{\text{[Blue]} + \text{[Red Hatched]} + \text{[Green Dashed]}}$$

**How it is computed:**

$$\text{Transfer Eff.} = \frac{T_{\text{ideal}}}{T}$$



“Ideal” simulation: Communications are instantaneous

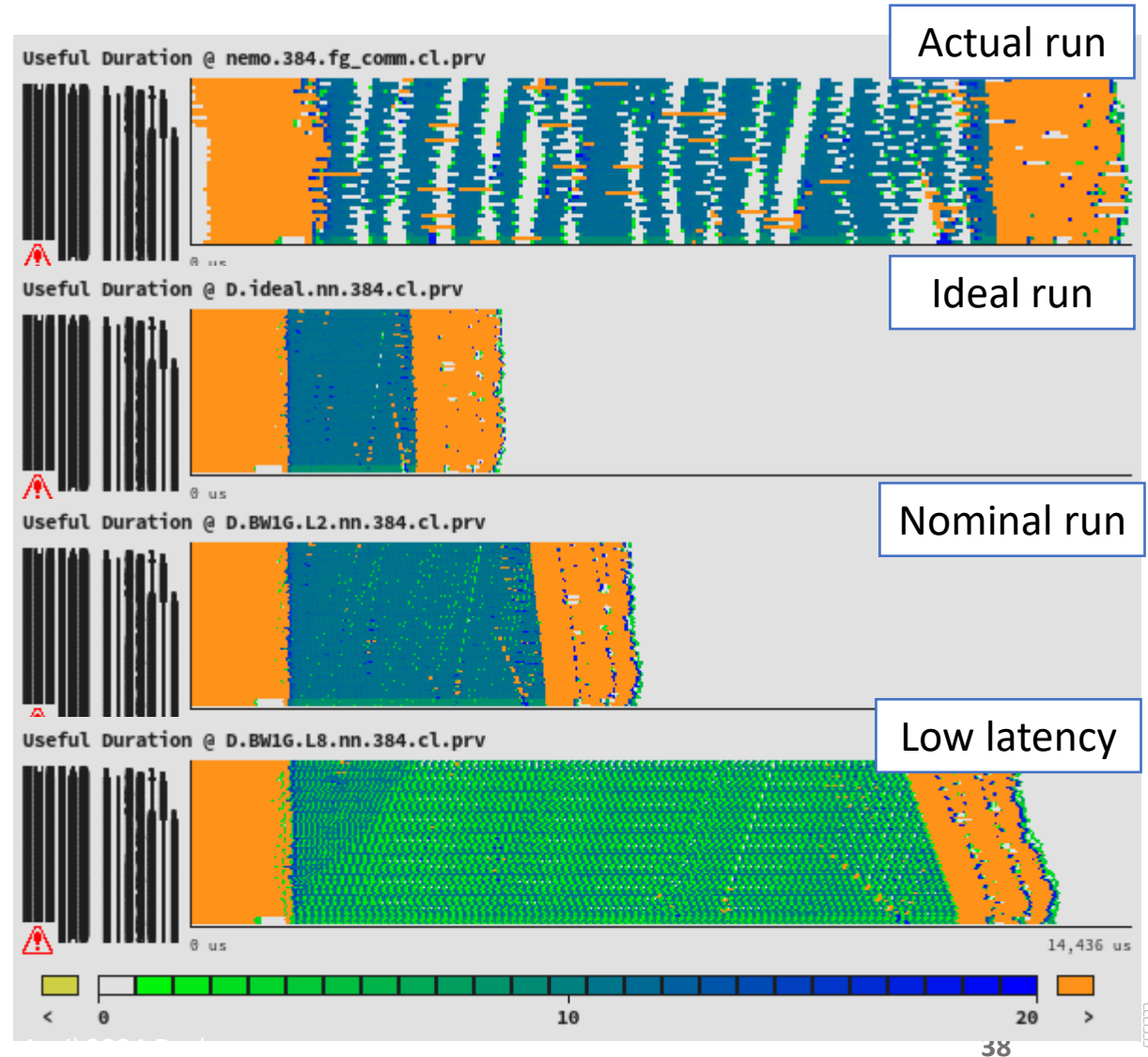


# Transfer Efficiency

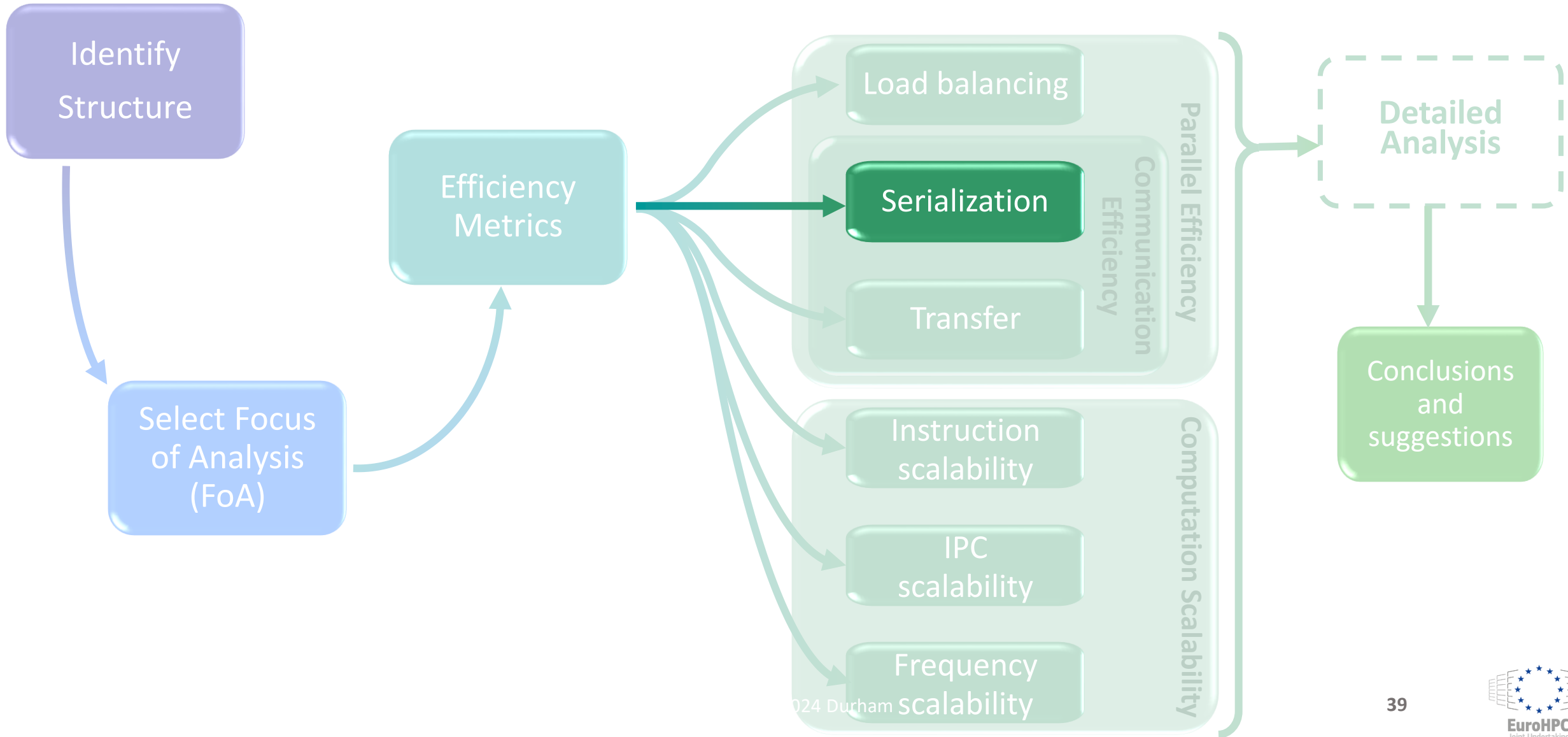


**Interpretation:** Low values indicate that the execution is suffering from a high overhead of the runtime or a poor latency or bandwidth of the network.

- What to look next?
  - Determine if the transfer problem is Bandwidth or Latency
    - Note that in Latency we include the overhead of the communication library
  - Use Simulations of Dimemas
    - Different BW and latency



# The journey



# Serialization Efficiency



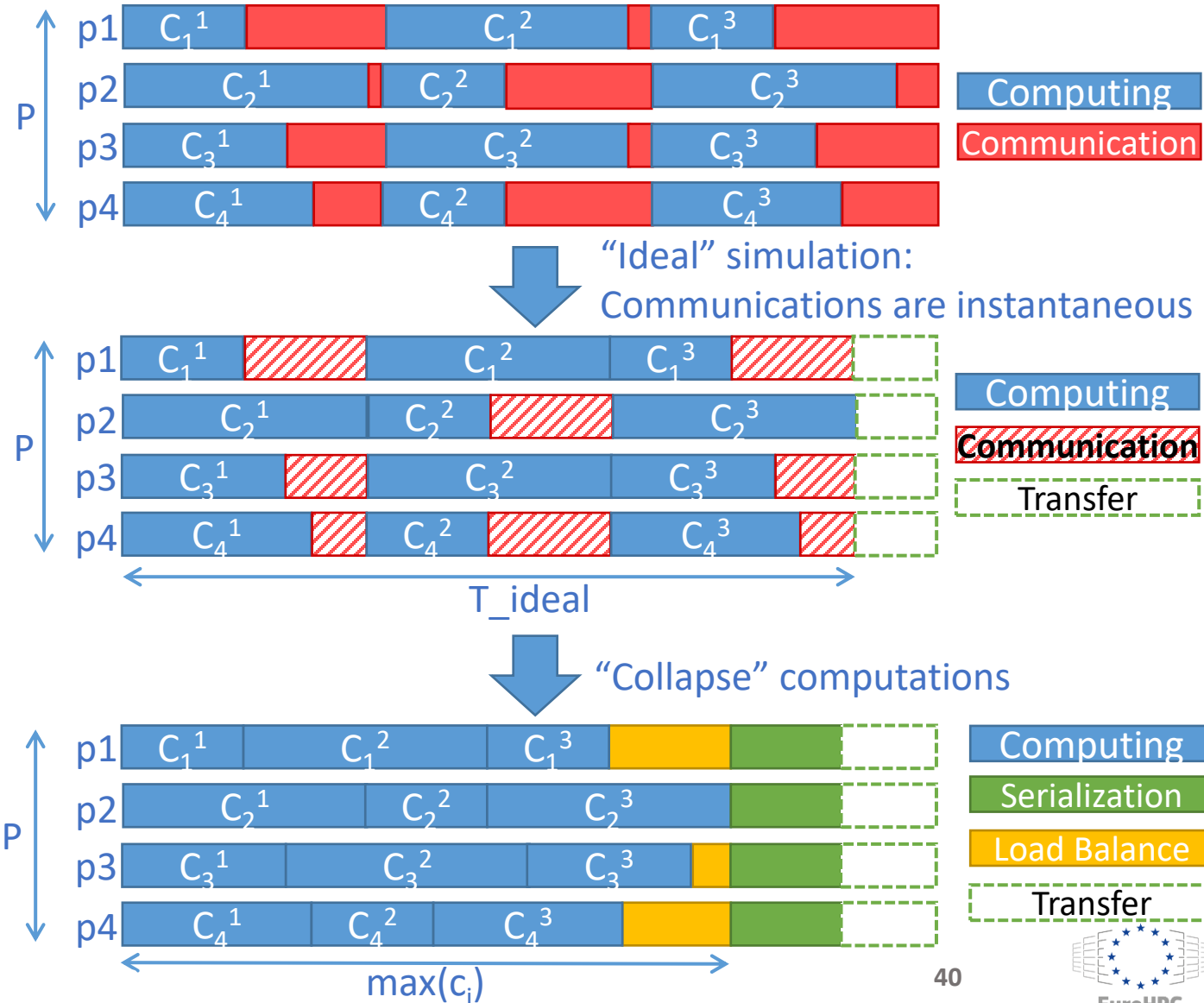
**Quantifies:** Inefficiencies caused by circular dependences or non-uniform imbalances

**How it is computed:** Ratio between useful computation of most loaded process and elapsed time in the ideal simulation

$$\text{Serialization Eff.} = \frac{(\text{blue} + \text{yellow})}{(\text{blue} + \text{red-hatched})}$$

**How it is computed:**

$$\text{Serialization Eff.} = \frac{\max_{i=1}^P (c_i)}{T_{ideal}}$$

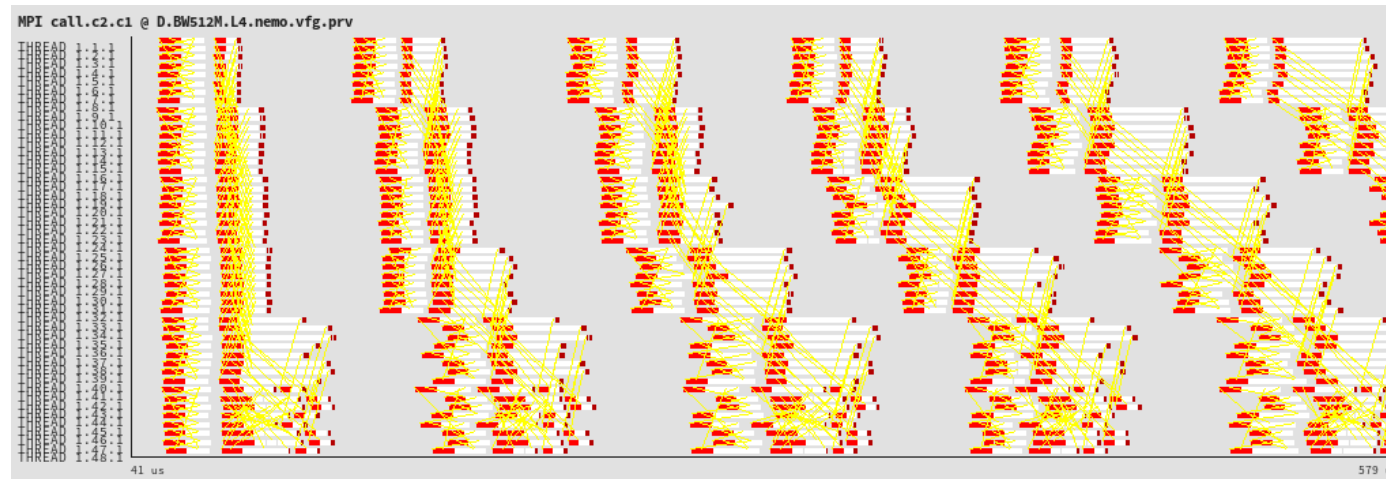


# Serialization Efficiency

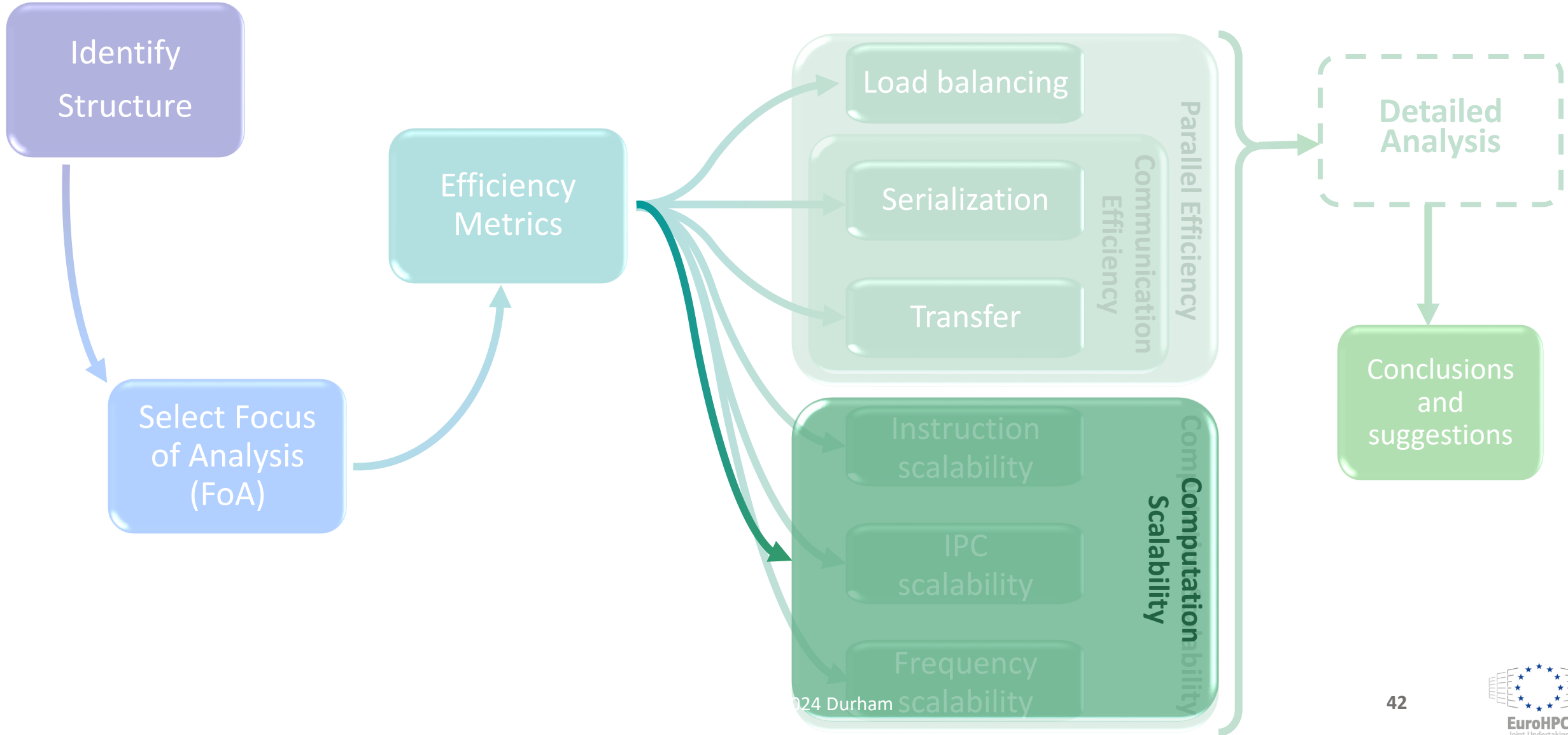


**Interpretation:** A low value indicates the existence of circular dependences. They can be caused by actual algorithmic serialization, irregularities in the load of processes during the execution or noise.

- What to look next?
  - Discard noise
    - Cycles per us
  - Try to identify causes for circular waits.
  - Understand semantics of the communication
    - MPI Calls



# The journey



# Computation Scalability



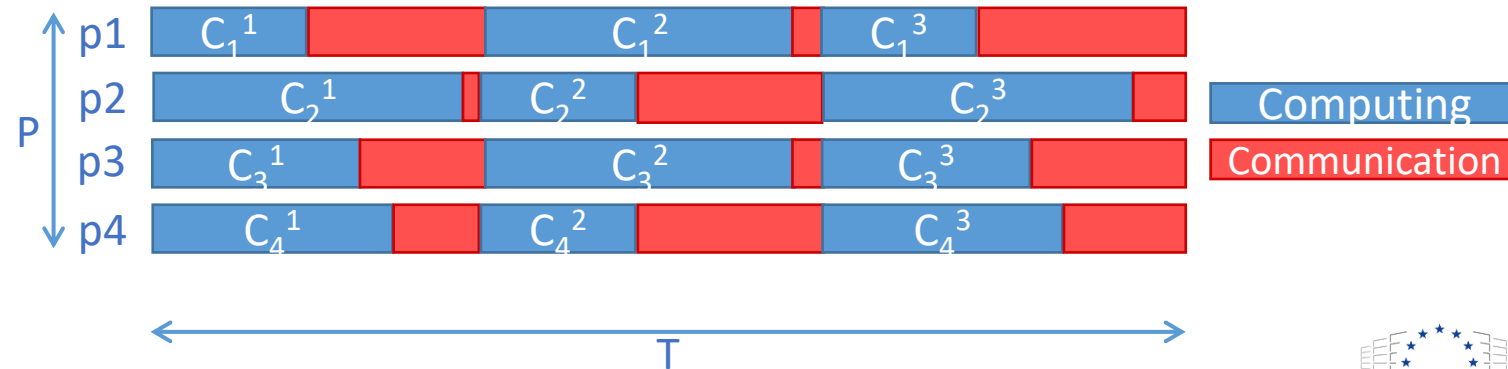
**Quantifies:** How the time spent computing scales with respect to the reference case.

**How it is computed:** Ratio between time spent on useful computation in the reference case and the time spent on useful computation on current run.

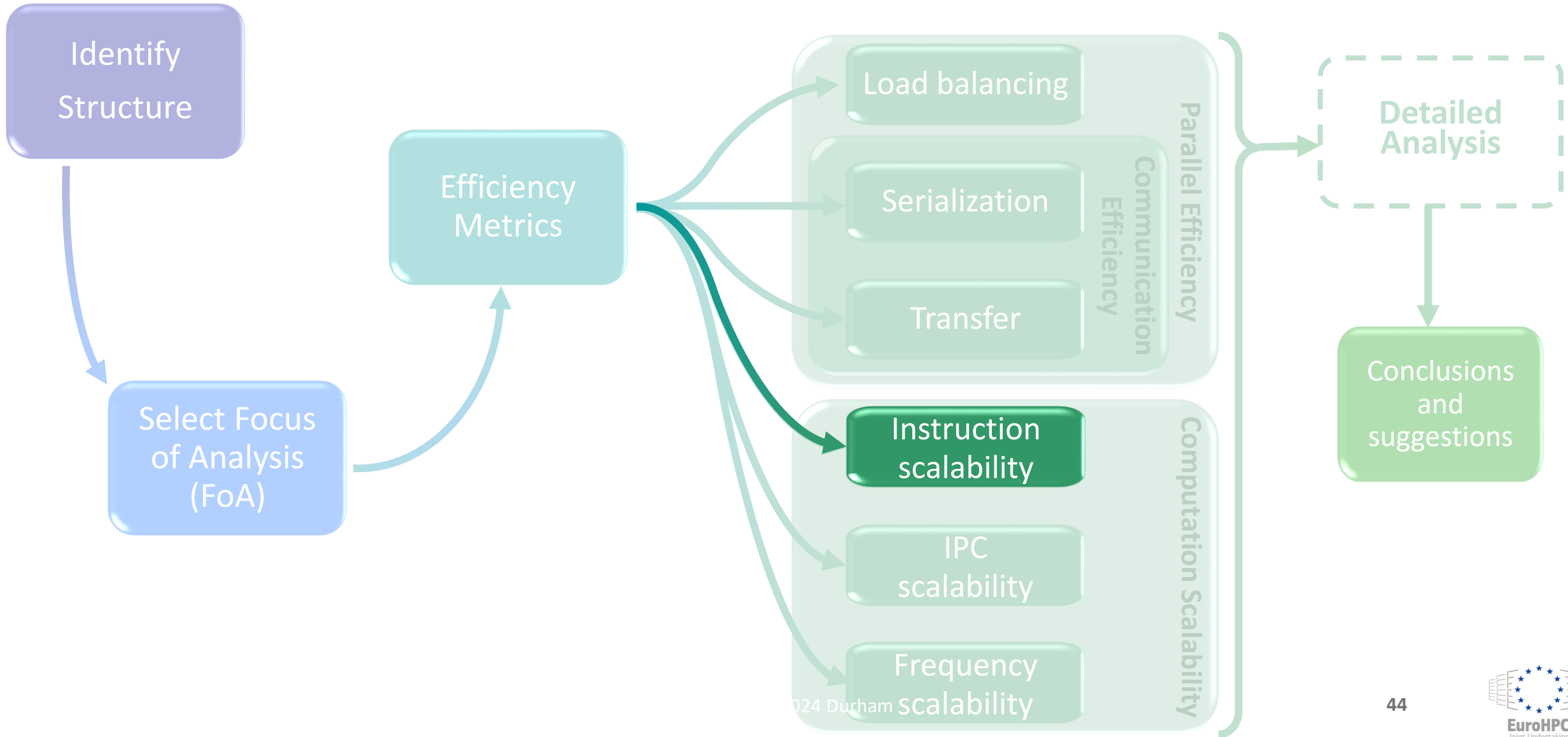
**How it is computed:**  
$$\text{Computation Sc.} = \frac{C_{ref}}{C_{current}}$$

**Interpretation:** A low value indicates time grows per core count. Ideally the total compute time to solve a problem should be constant, independent of core count (in strong scaling).

- Relative metric, based on a reference case.
  - $C_{ref}$  = Useful Computation of the reference run
  - $C_{current}$  = Useful computation current run
- 100% for the reference case
- Can be weak or strong scaling
  - Strong scaling for all the formulas presented
    - In strong scaling we assume total compute time and total number of instructions should remain constant as we increase the number of processes
  - In an analogous way weak scaling can be computed
- High level metric composed by three child metrics based on:
  - $T = \# \text{ instr.} / (\text{IPC} * \text{freq})$



# The journey



# Instructions Scalability



**Quantifies:** How the number of instructions scales with respect to the reference case.

**How it is computed:** Ratio between number of instructions executed on the reference case and the number of instructions executed on current run.

**How it is computed:**

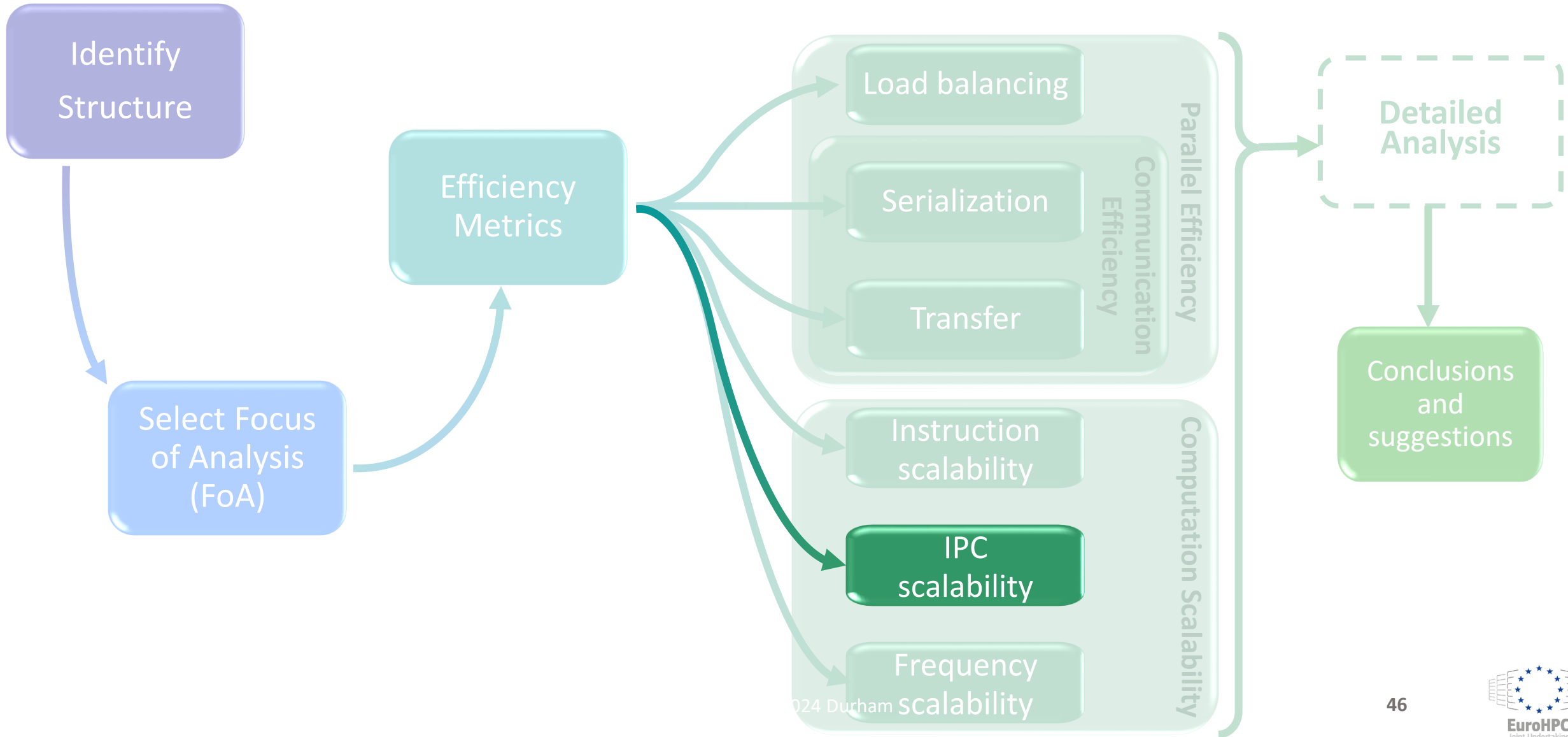
$$\text{Instructions Sc.} = \frac{I_{ref}}{I_{current}}$$

- Relative metric, based on a reference case.
  - $I_{ref}$  = Total number of instructions executed on the reference run
  - $I_{current}$  = Total number of instructions executed on the current run
- 100% for reference case

**Interpretation:** A value less than 100 indicate that the total number of instructions to solve the problem grows with core count, which ideally should not be the case.

May be caused by code replication (computed by all processes, by an increase in the surface to volume ratio when computations on the surface/boundary are "replicated", ...

# The journey





**Quantifies:** How the IPC scales with respect to the reference case.

**How it is computed:** Ratio between average IPC on the current run and the average IPC on the reference run.

**How it is computed:**

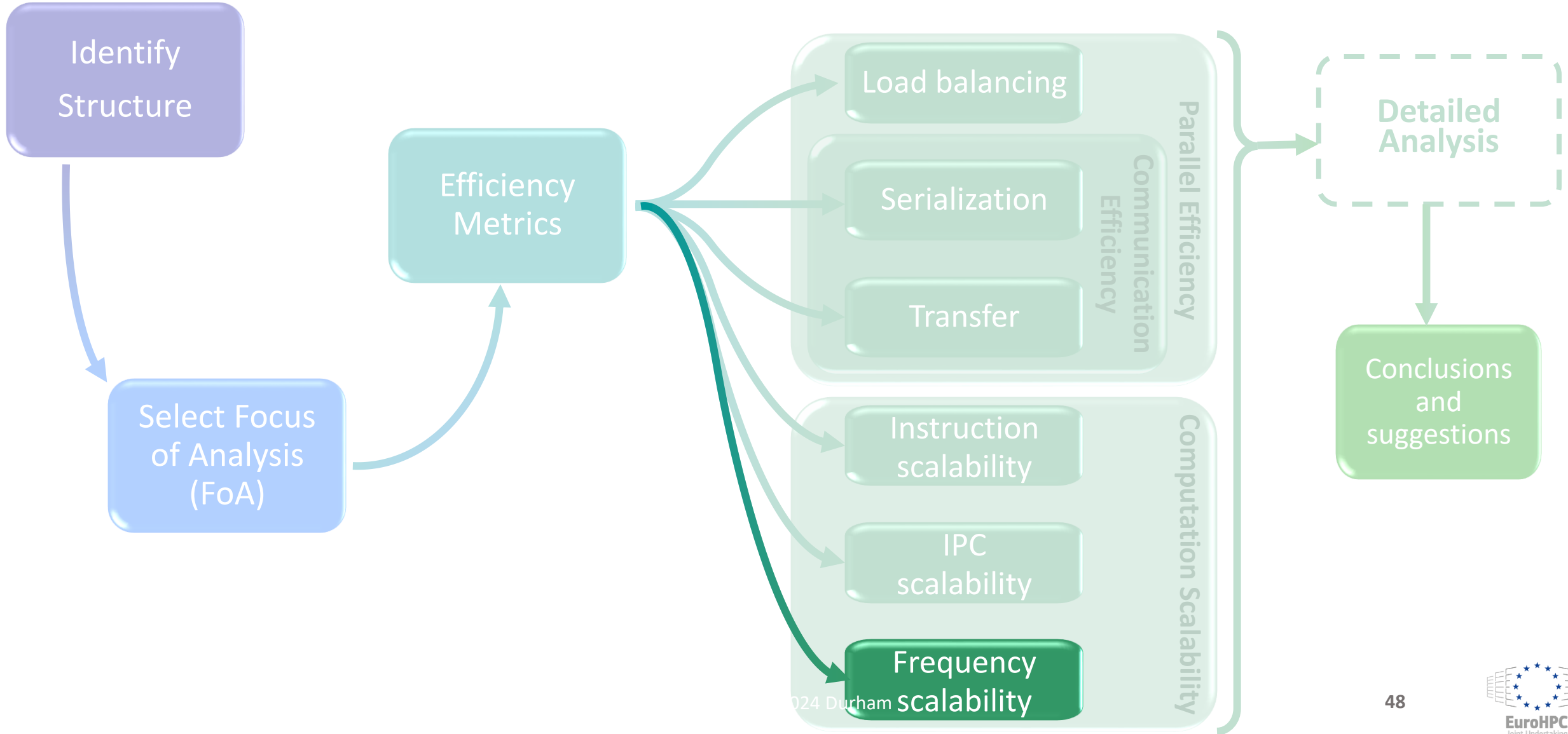
$$IPC_{Sc.} = \frac{IPC_{current}}{IPC_{ref}}$$

- Relative metric, based on a reference case.
  - $IPC_{ref}$  = Average IPC of the reference run
  - $IPC_{current}$  = Average IPC current run
- 100% otherwise

**Interpretation:** A value less than 100 indicates that the IPC for the specific core count is worse than that of the reference case. May be caused by different locality behavior, contention on resources such as memory bandwidth, ....

A value above 100 indicates a higher IPC than the reference case this can be produced by cache effects for example.

# The journey



# Frequency Scalability



**Quantifies:** How the frequency scales with respect to the reference case.

**How it is computed:** Ratio between average frequency on the current run and the average frequency on the reference run.

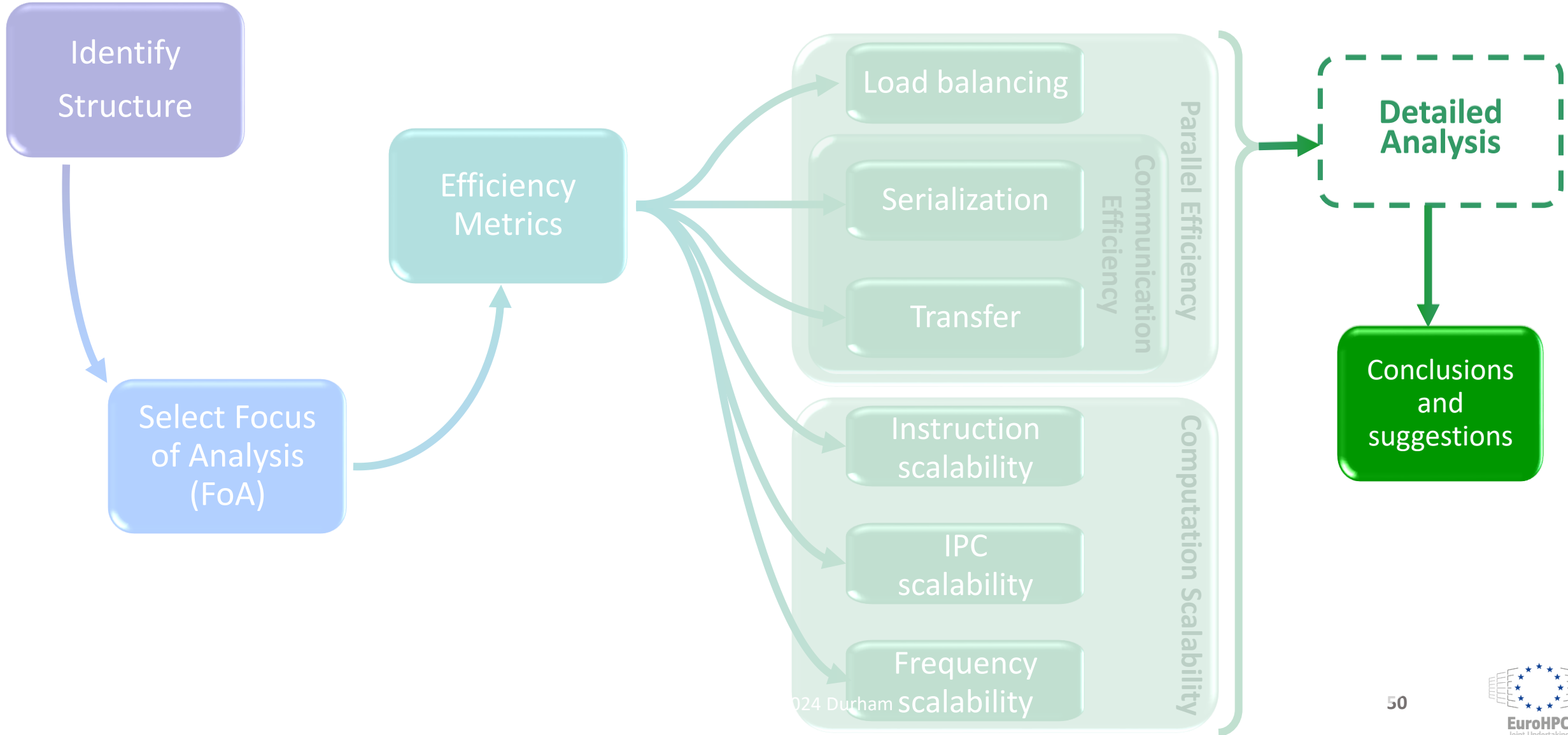
**How it is computed:**

$$\text{Frequency Sc.} = \frac{\text{Freq}_{\text{current}}}{\text{Freq}_{\text{ref}}}$$

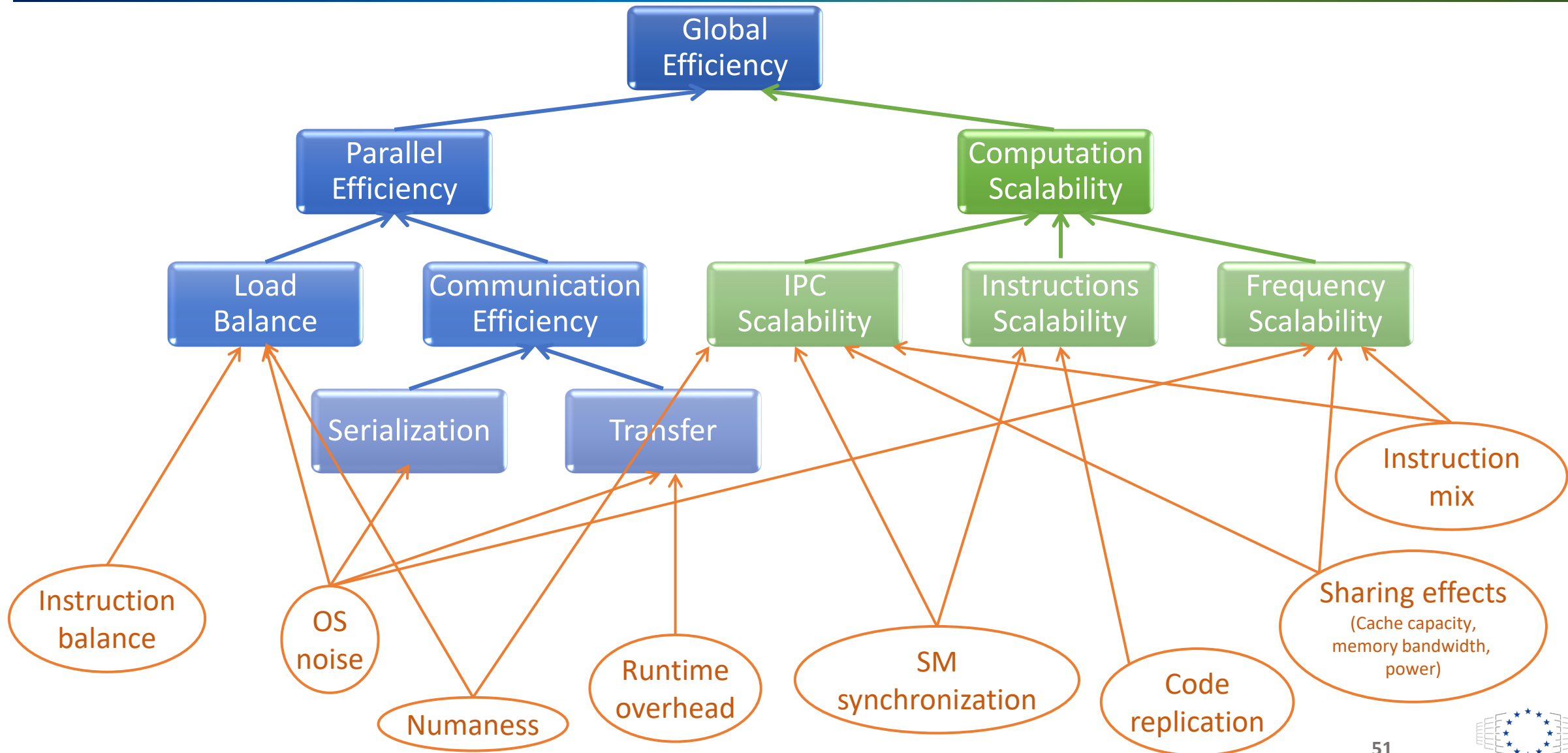
- Relative metric, based on a reference case.
  - $\text{Freq}_{\text{ref}}$  = Average frequency on the reference run
  - $\text{Freq}_{\text{current}}$  = Average frequency on the current run
- 100% for the reference case

**Interpretation:** A value less than 100 indicates that the frequency is lower than the reference frequency. This may be caused by "preemptions", power management measures, ...

# The journey



# Factors and causes





# Take away messages

# Take away messages



- **POP Services:** **Free** and available for all European HPC users or developers
  - Request your **performance assessment** service here:
    - <https://www.pop-coe.eu/request-service-form>
- **POP Methodology:** Agnostic to the tools, gives hints on main efficiency bottlenecks
  - Covers: MPI, OpenMP, hybrid, MPI+GPUs...
- **POP Tools:** Available to be used, will be presented in the following days
- **POP Training:** We also do advanced trainings, covering more tools and more functionalities, ask us if you are interested



# Performance Optimisation and Productivity 3

A Centre of Excellence in HPC

## Contact:

 <https://www.pop-coe.eu>

 [pop@bsc.es](mailto:pop@bsc.es)

 [@POP\\_HPC](#)

 [youtube.com/POPHPC](https://www.youtube.com/POPHPC)

