XXI Jornada Técnica de AdaSpain
17 de Mayo de 2024

Modelado de GPUs en MAST.

Iosu Gomez[a,b]

Juan M. Rivas[a]

J. Javier Gutiérrez[a]

Jorge Parra [b]

Unai Díaz de Cerio [b]

[a] Universidad de Cantabria
[b] Ikerlan Research Centre

Funded by MCIN/ AEI /10.13039/501100011033/ FEDER "Una manera de hacer Europa" under grants PID2021-124502OB-C42 and PID2021-124502OB-C44 (PRESECREL).

Índice

1. Motivación y objetivos
2. Contexto
3. Análisis de tiempos de respuesta
4. Extendiendo MAST-2
5. Marco de optimización

1.

Motivación y objetivos.

2. Contexto.
3. Análisis de tiempos de respuesta.
4. Extendiendo MAST-2.
5. Marco de optimización.

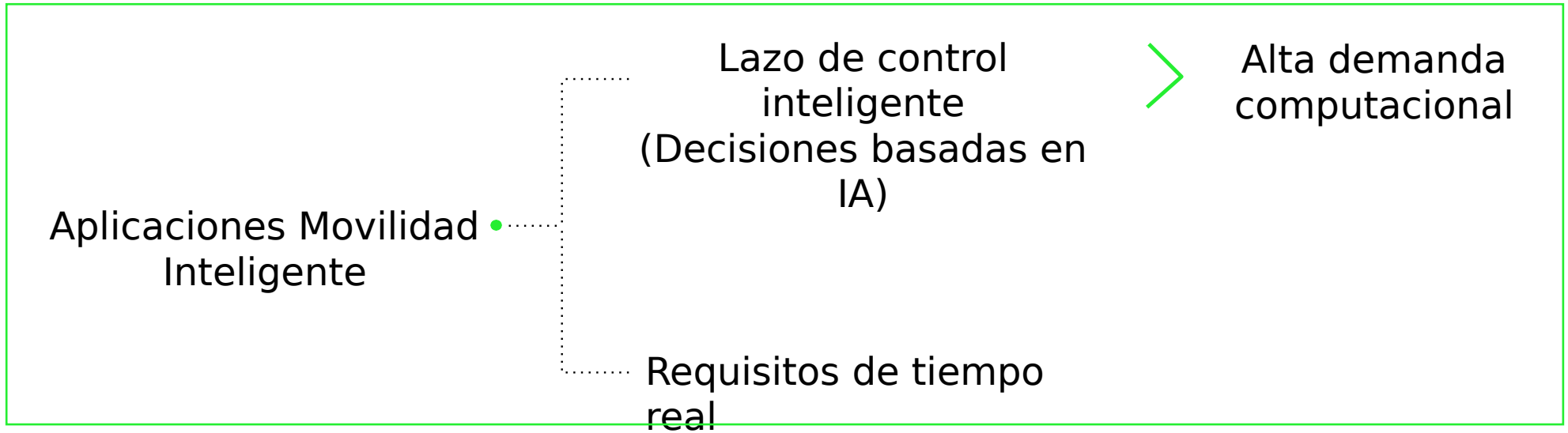
ISTR- UC UNIVERSIDAD DE CANTABRIA

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Motivación.

En aplicaciones de **Movilidad Inteligente**, las tareas de control requieren de una alta computación mientras se mantienen requisitos de **tiempo real**.

Esto lleva a utilizar nuevas **plataformas heterogéneas** con aceleradores hardware como **GPUs**.



Objetivos.

- Implementar una metodología para incorporar las GPUs a los sistemas de tiempo real
- Utilizar y extender técnicas y herramientas existentes y validadas que se presentan en MAST
- Tres etapas:

Análisis de tiempos de respuesta → **Extender el metamodelo MAST** → **Framework de optimización**

2.

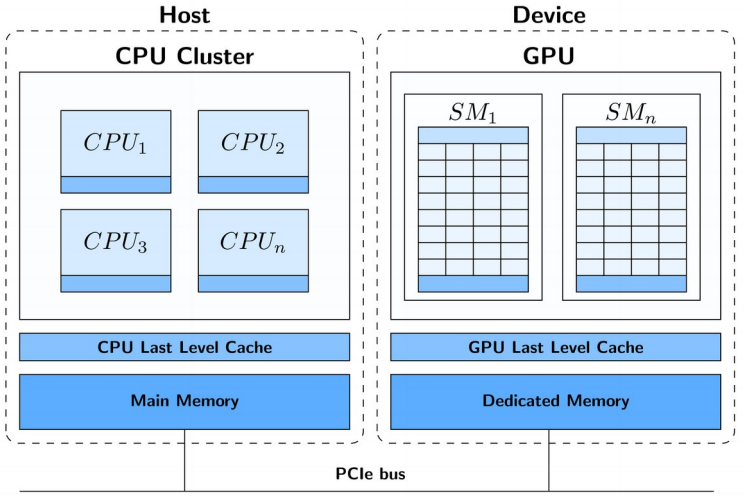
Contexto.

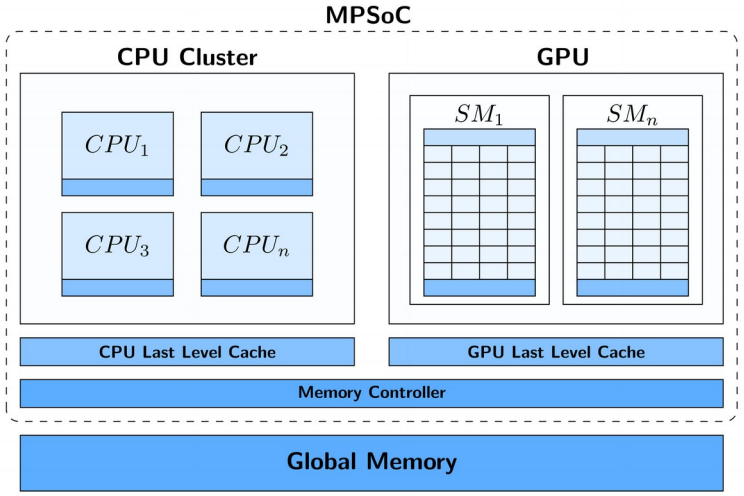
3. Análisis de tiempos de respuesta.
4. Extendiendo MAST-2.
5. Marco de optimización.

ISTR
UC UNIVERSIDAD DE CANTABRIA
ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Graphics Processing Unit (GPU

- Coprocesadores para acelerar la carga de trabajo en la CPU gracias a su gran capacidad de paralelización.
- Dos tipos de arquitecturas:
  - Discretas**:** conectadas normalmente mediante bus PCIe a la placa base del sistema.
  - **Integradas:** embebidas dentro de un MPSoC compartiendo la memoria del sistema.
- Mecanismos internos de planificación desconocidos, impredecibilidad de los tiempos de ejecución.

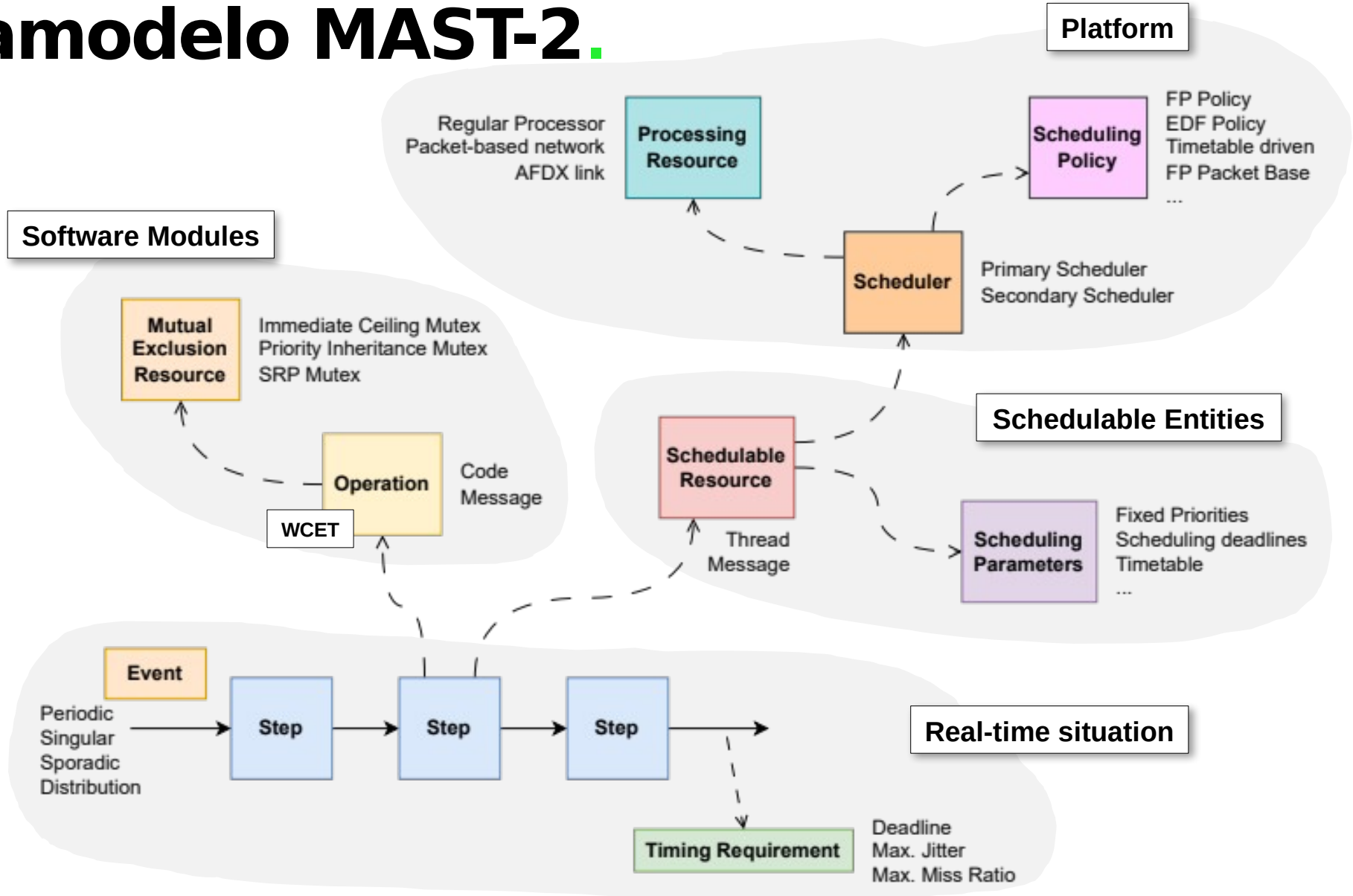




MAST-2.

- Define el modelo para describir el comportamiento temporal de sistemas de tiempo real.
- Alineado con el estándar **OMG MARTE**.
- Contiene herramientas para diferentes técnicas de análisis.
- Disponible en: https://mast.unican.es/.

Metamodelo MAST-2.



3.

Análisis de tiempos de respuesta.

4. Extendiendo MAST-2.
5. Marco de optimización.

ISTR- UC UNIVERSIDAD DE CANTABRIA

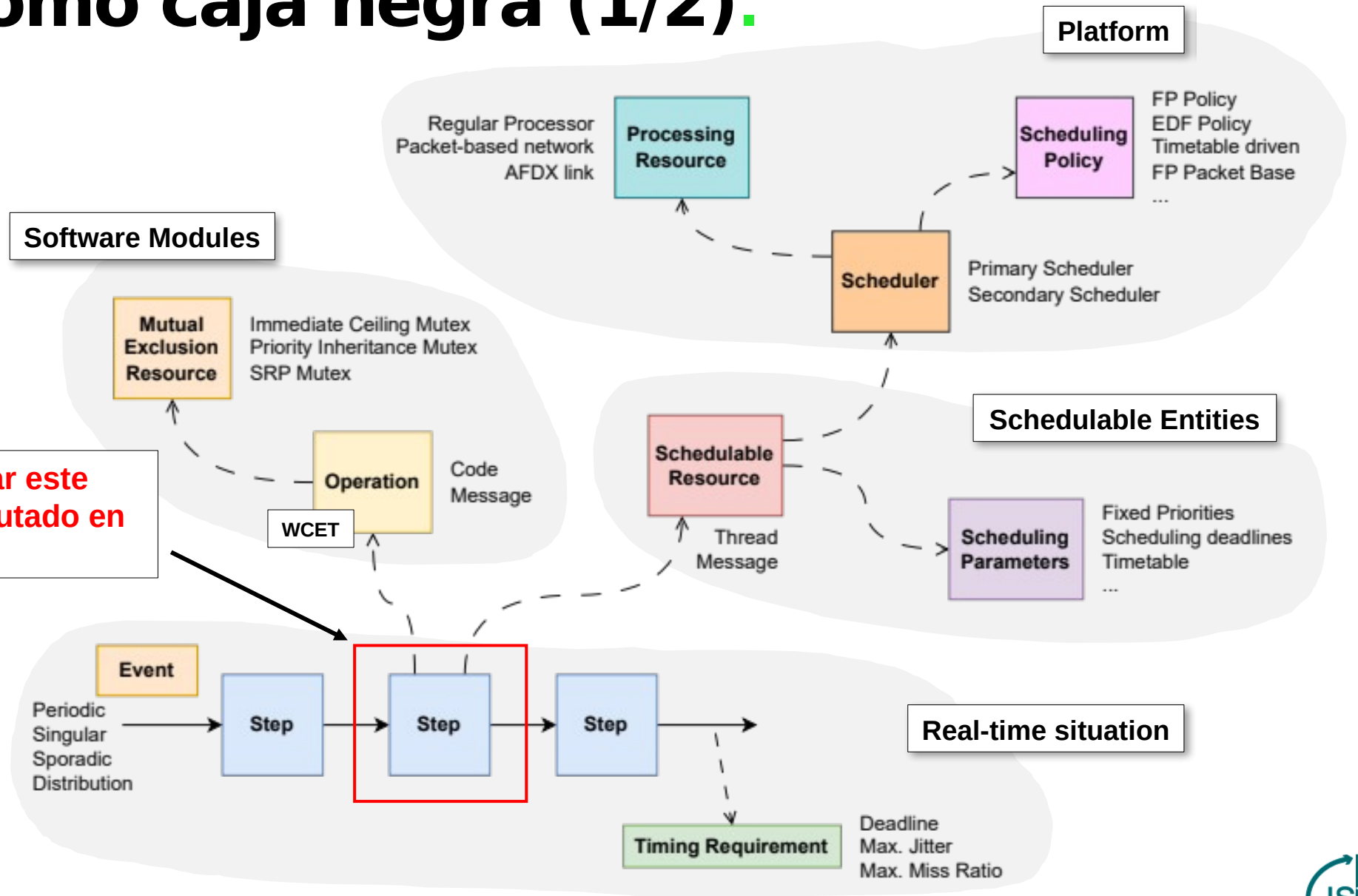
ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Modelado de GPUs.

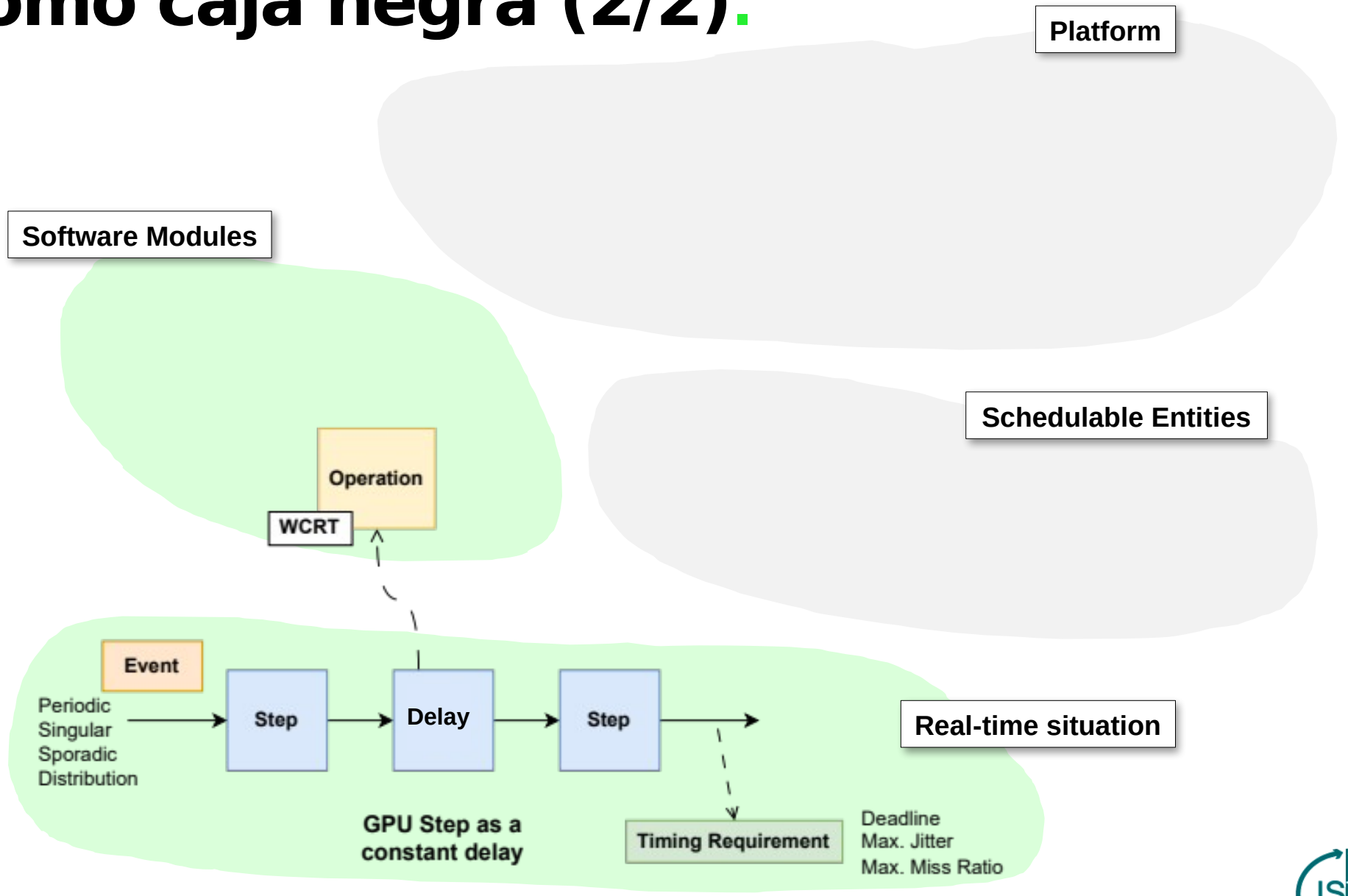
- No se analiza el funcionamiento interno de la GPU
- Se miden los tiempos de respuesta de la GPU

- Describe componentes de la GPU
- Meta-modelo preparado para futuro
- Se desconoce cómo explotar en este momento

GPU como caja negra (1/2).



GPU como caja negra (2/2).



GPU como caja negra (2/2).

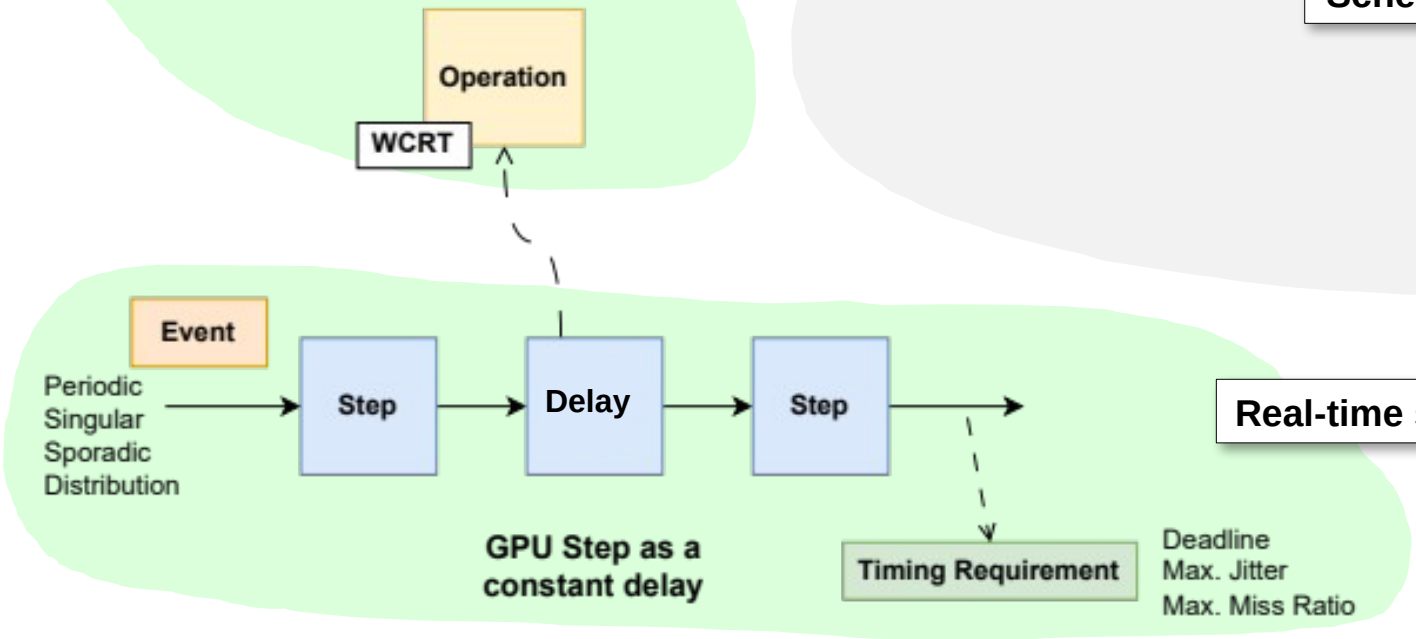
Platform

Analizable con diferentes técnicas de análisis para sistemas distribuidos:

- Diferentes variantes de técnicas basadas en ***offsets***
- Palencia and González (1999), Redell (2003), Maki-Turja and Nolin (2008)

Software Modules

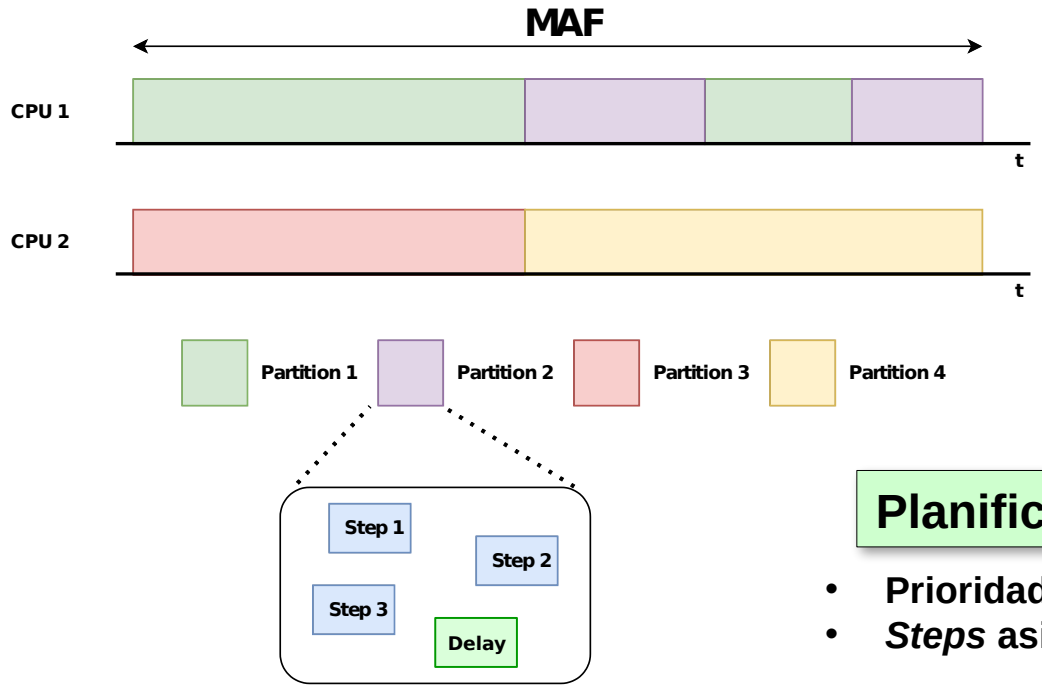
Schedulable Entities



Real-time situation

Análisis de tiempos de respuesta (1/2).

- Se utiliza el **particionado temporal** para controlar:
  - Interferencias CPU-GPU
  - Acceso concurrente a la GPU



Planificador primario

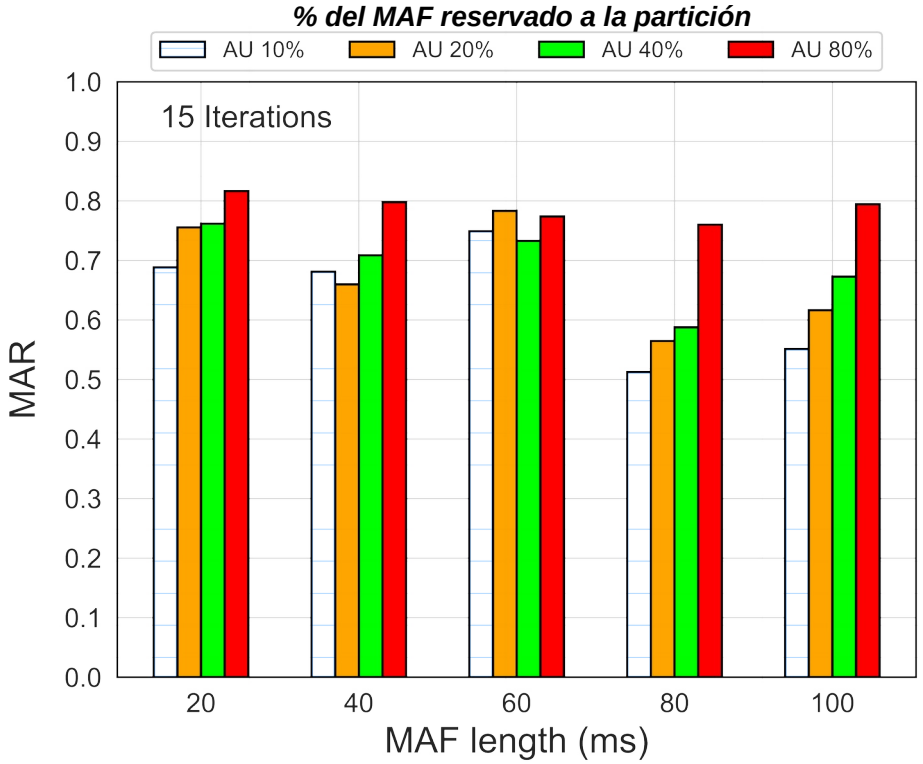
- **Planificación *table-driven***
- **Particiones temporales asignadas a procesadores**

Planificador secundario

- **Prioridades fijas dentro de la partición**
- ***Steps* asignados a particiones**

Análisis de tiempos de respuesta (2/2)

- **Evaluación** sobre caso de uso industrial:
  - Optimizador de perfil de vía de CAF
  - Grandes cargas de trabajo descargadas en la GPU
  - Se ha evaluado el pesimismo del análisis



Comparación de los WCRT analíticos y medidos

$$MAR = \frac{Measurement}{Analytical}\ Ratio$$

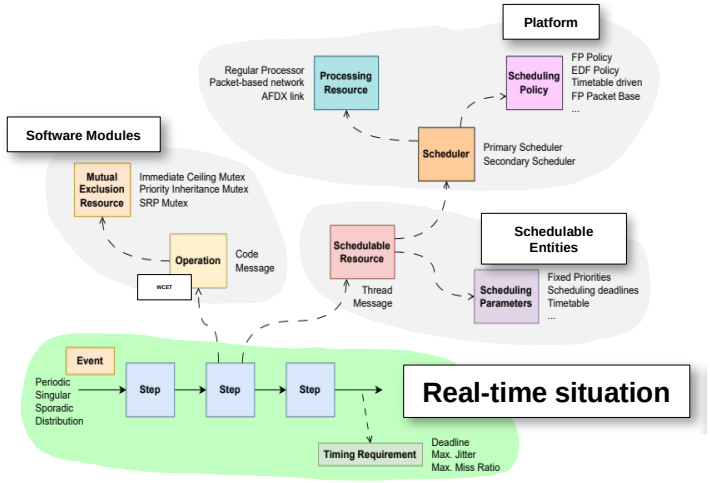
4.

Extendiendo MAST-2.

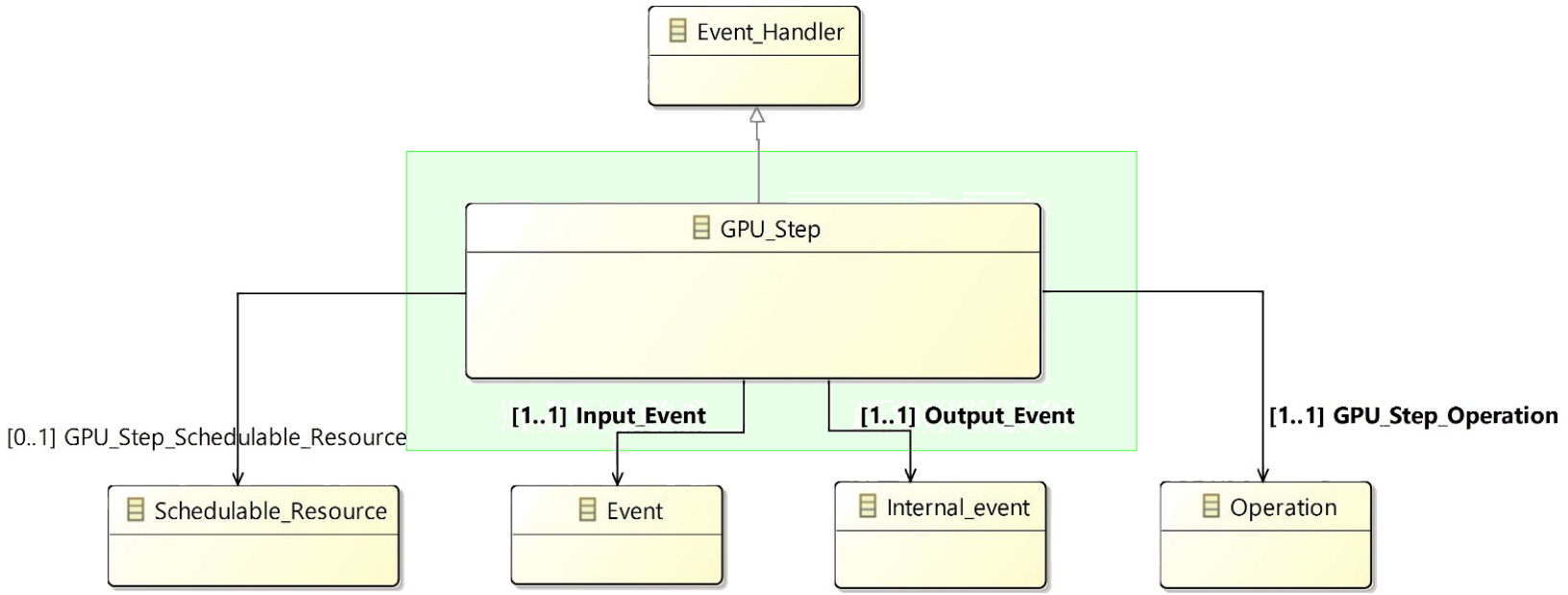
5. Marco de optimización.

ISTR UC UNIVERSIDAD DE CANTABRIA

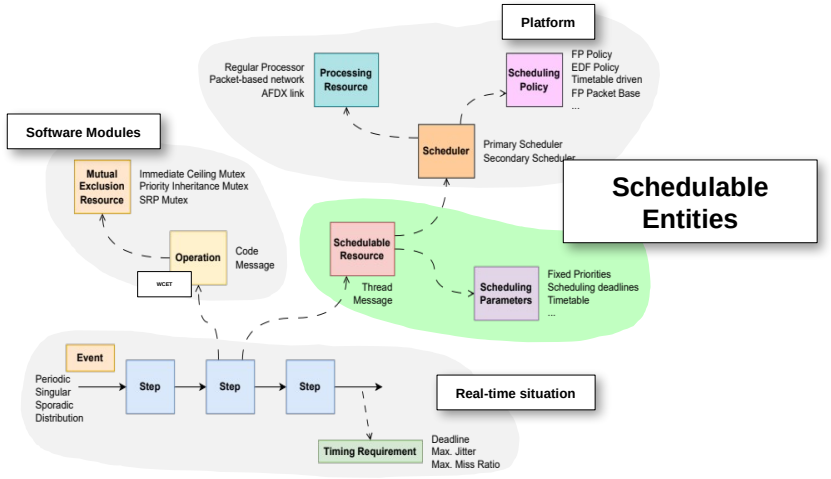
ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

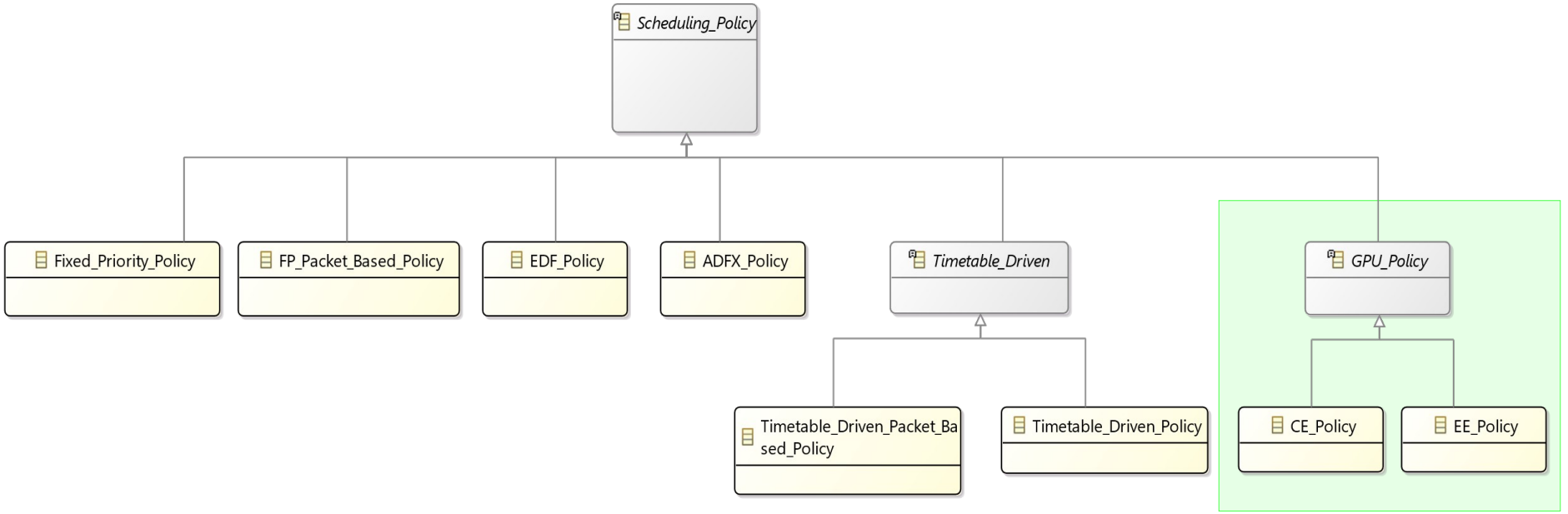


Descripción detallada (1/3).

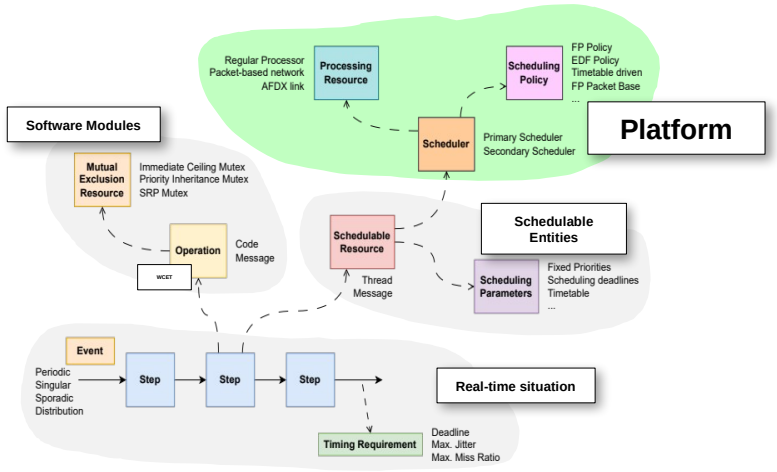


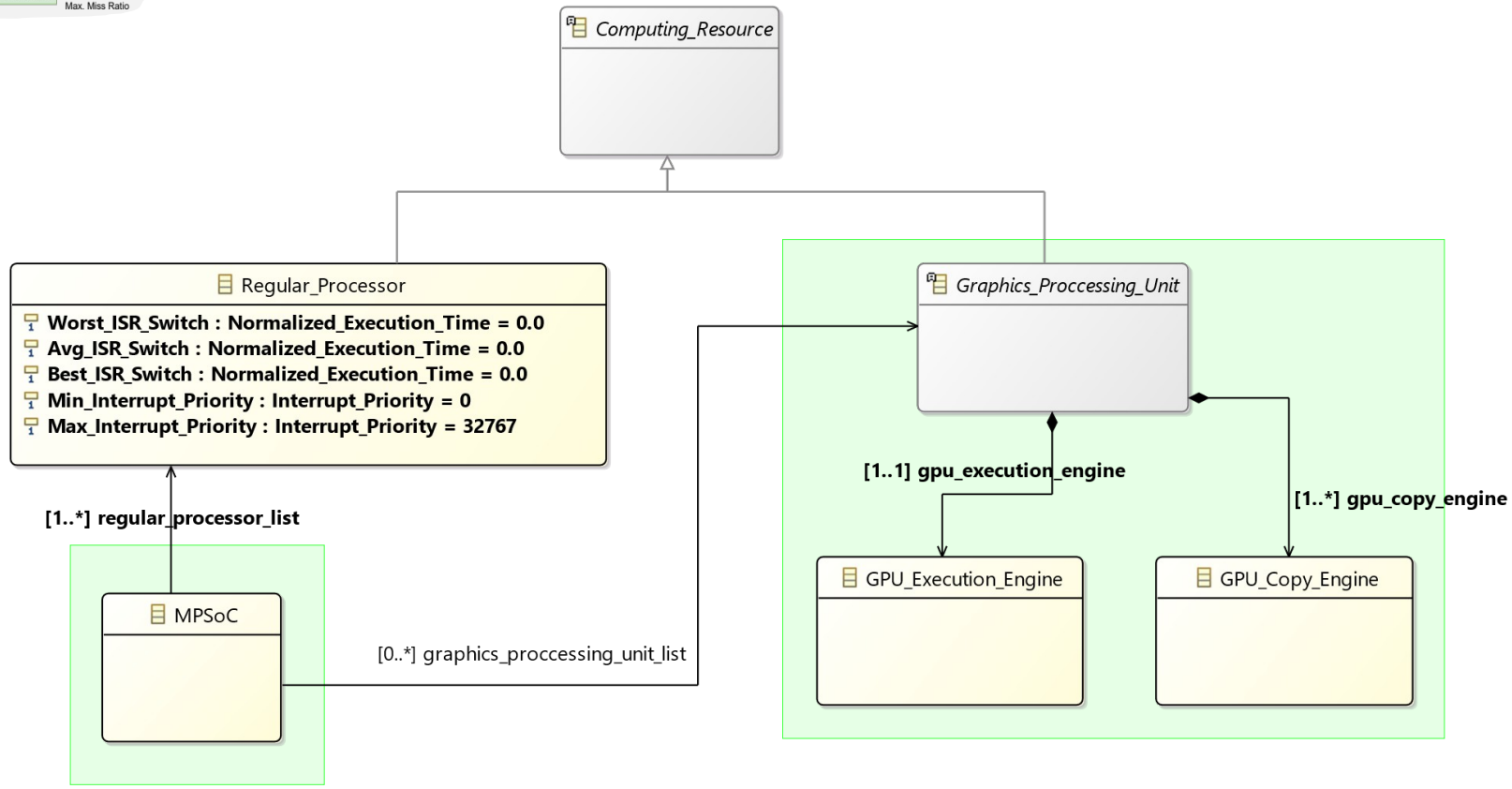
Descripción detallada (2/3).





Descripción detallada (3/3).





5.

Marco de optimización

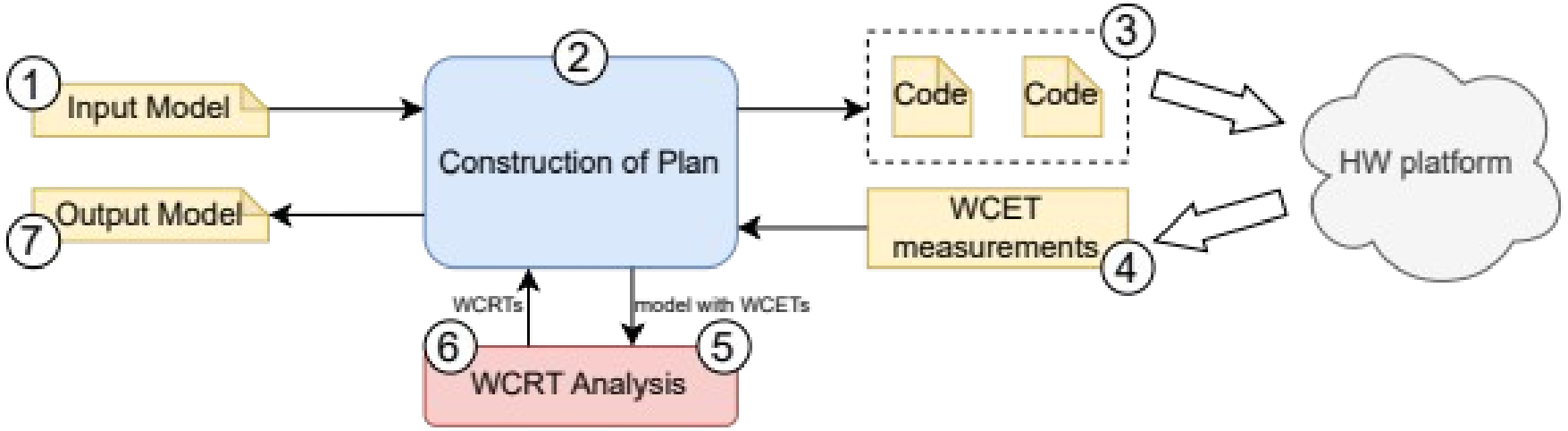
ISTR UC UNIVERSIDAD DE CANTABRIA

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Optimización (work in progress).

- Hay muchos parámetros que influyen en la planificabilidad del sistema:
  - Particiones, asignación de tareas y particiones, prioridades …
  - Interdependencias: interferencia en varios niveles de memoria que aumentan los WCET.

- Solución propuesta:

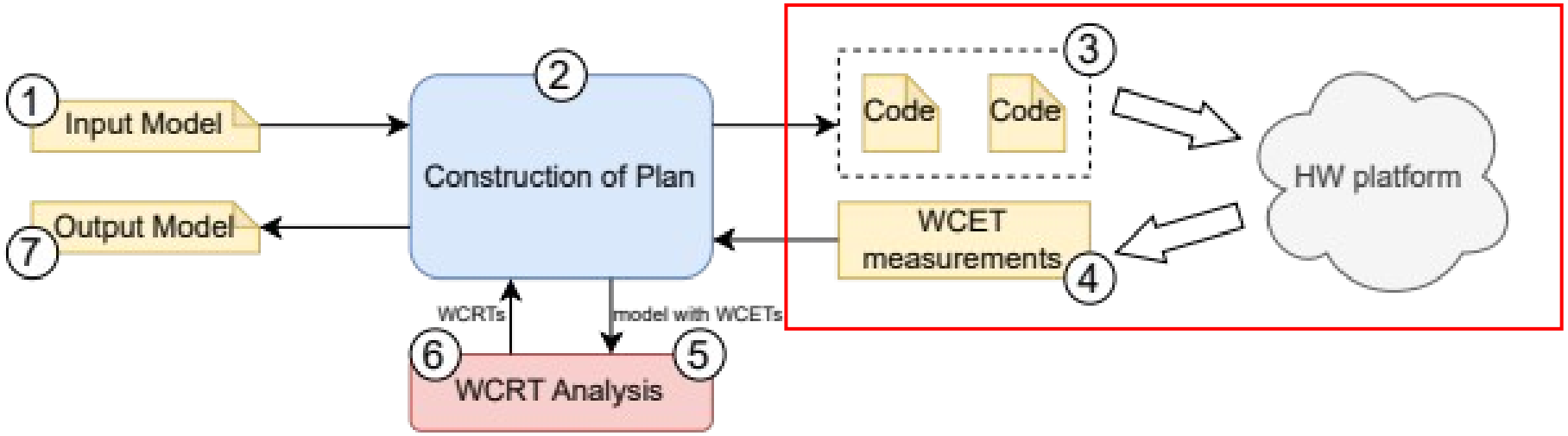


Optimización (work in progress).

- Hay muchos parámetros que influyen en la planificabilidad del sistema:
  - Particiones, asignación de tareas y particiones, prioridades …
  - Interdependencias: interferencia en varios niveles de memo
    que aumentan los WCET.

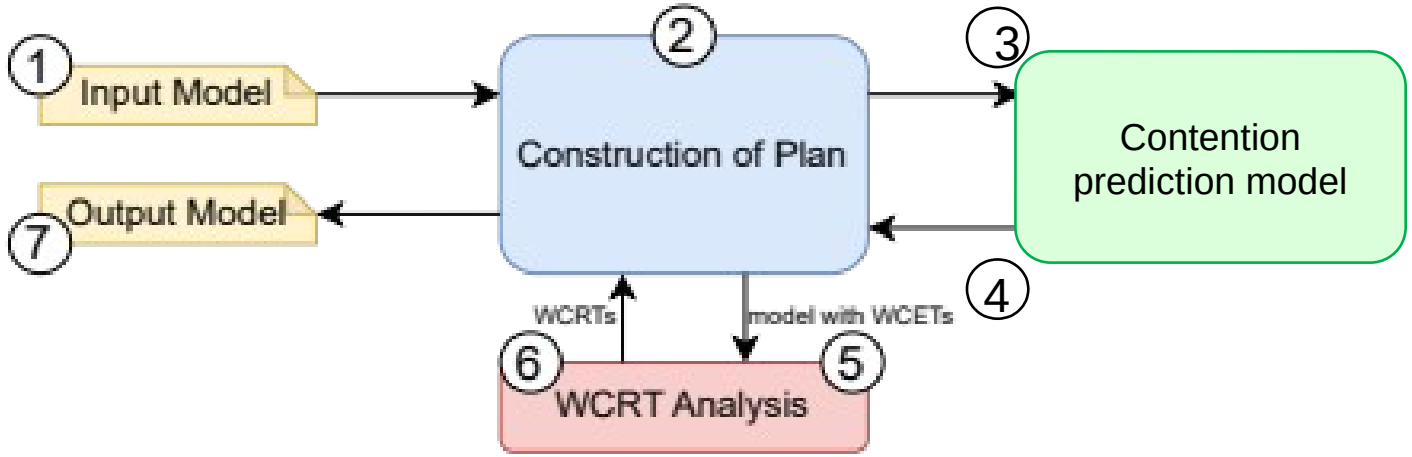
- Solución propuesta:

Método muy lento. Para mejorar se podían implementar técnicas de estimación.



- Modelo que predice la contención en la memoria.
  - **Input:** *Performance counters* de las tareas medidas de manera aislada.
  - **Output:** Tiempo máximo por contención

- A partir de estas predicciones se calculan los nuevos WCETs



¡Gracias por su atención!

Iosu Gomez Iturrioz
iosu.gomez@ikerlan.es

ISTR UC UNIVERSIDAD DE CANTABRIA

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE