



Multiprocessor Real-Time Scheduling with **Hierarchical Processor Affinities**

Vincenzo Bonifaci
IASI-CNR

Björn Brandenburg
MPI-SWS

Gianlorenzo D'Angelo
Gran Sasso Science Institute

Alberto Marchetti-Spaccamela
Sapienza Università di Roma



Max
Planck
Institute
for
Software Systems



This Paper

- Real-time scheduling with **restricted processor affinities** (*each task may run only on certain processors*)
- Identify ***hierarchical (or laminar) affinities*** as a special case of great practical relevance
- Non-obvious **online scheduling algorithm** with **improved runtime complexity**
- Performance characterization:
 1. **speed-up** bound in case of *clustered* or *bi-level* affinities
 2. prototype **implementation in LITMUS^{RT}** and **overhead evaluation** on 24-core Xeon multicore platform

Background

Processor Affinity

- interface to ***restrict the set of processors*** on which a task may be scheduled
- widely available in multiprocessor (real-time) OSs

Linux: **`sched_setaffinity()`**

FreeBSD: **`cpuset_setaffinity()`**

Windows: **`SetThreadAffinityMask()`**

QNX: **`ThreadCtl(_NTO_TCTL_RUNMASK)`**

VxWorks: **`taskCpuAffinitySet()`**

Arbitrary Processor Affinity (APA) Scheduling (*Gujarati et al., 2013*)

- first analysis of processor affinity in real-time systems
- the usual sporadic task model: $\mathbf{C}_i, \mathbf{D}_i, \mathbf{T}_i$
- set of (identical) processors $\mathbf{\Pi}_1 \dots \mathbf{\Pi}_m$
- plus an ***arbitrary per-task affinity set***

$$\alpha_i \subseteq \{ \mathbf{\Pi}_1, \dots, \mathbf{\Pi}_m \}$$

Strong vs. Weak APA Scheduling

(*Gujarati et al., 2014*)

weak APA invariant

a job is **backlogged** only if all processors in its affinity execute jobs of **equal or higher priority**

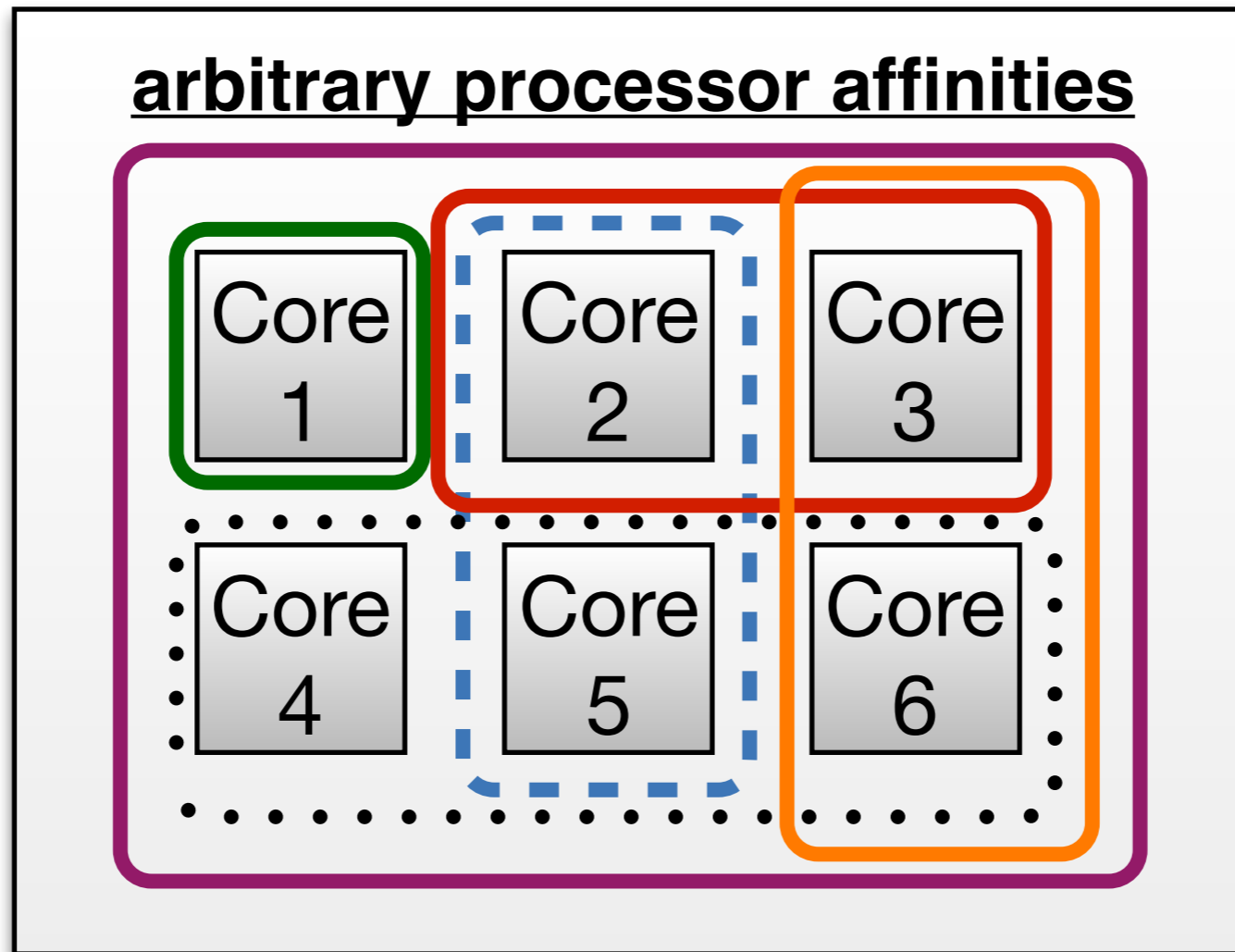
- Linux, QNX, etc.
- easier to implement

strong APA invariant

weak invariant + **no way to “re-arrange” higher-priority jobs** to free up a core for a backlogged job

- better schedulability
- this paper

Arbitrary Affinities: Difficult Scheduling Problem



- difficult to analyze
- difficult to schedule at runtime

Basic Operations

Job Arrival: preemption necessary?

- for each core in affinity, check if new job can be placed
- **weak APA:** only by preempting lower-priority tasks
- **strong APA:** possibly by *shifting* higher-priority tasks to other cores

Job Departure: schedule backlogged job?

- for each backlogged job, check if freed processor can be used
- **weak APA:** only if freed processor is in affinity set
- **strong APA:** possibly by *shifting* higher-priority tasks to other cores

Prior Strong APA Scheduling Results

	Strong APA (<i>Gujarati et al., 2014</i>)
Job arrival cost	$O(m^2)$
Job departure cost	$O(nm)$
Speed-up bound	—
Implemented in OS?	—
Schedulability test	sufficient

Difficult to improve
the general case.
(combinatorial
structure)

But what if we rule
out pathological
combinations?

n ...number of tasks

m ...number of cores

Hierarchical Processor Affinities (HPA)

Arbitrary Processor Affinities?

Why do users typically restrict processor affinities?

Arbitrary Processor Affinities?

Why do users typically restrict processor affinities?

- **cache affinity**: e.g., stay on same core / pair of cores / socket to maintain L1 / L2 / L3 affinity, respectively

Arbitrary Processor Affinities?

Why do users typically restrict processor affinities?

- **cache affinity:** e.g., stay on same core / pair of cores / socket to maintain L1 / L2 / L3 affinity, respectively
- **interrupt throughput:** e.g., execute network-facing daemon on same core that processes network interrupts

Arbitrary Processor Affinities?

Why do users typically restrict processor affinities?

- **cache affinity:** e.g., stay on same core / pair of cores / socket to maintain L1 / L2 / L3 affinity, respectively
- **interrupt throughput:** e.g., execute network-facing daemon on same core that processes network interrupts
- **interrupt isolation:** e.g., execute only on half of cores that do not handle device interrupts

Arbitrary Processor Affinities?

Why do users typically restrict processor affinities?

- **cache affinity:** e.g., stay on same core / pair of cores / socket to maintain L1 / L2 / L3 affinity, respectively
- **interrupt throughput:** e.g., execute network-facing daemon on same core that processes network interrupts
- **interrupt isolation:** e.g., execute only on half of cores that do not handle device interrupts
- **security isolation:** e.g., avoid micro-architectural timing channels by forcing sensitive and less trusted tasks to run on separate cores

Arbitrary Processor Affinities?

Why do users typically restrict processor affinities?

- **cache affinity**: e.g., stay on same core / pair of cores / socket to maintain L1 / L2 / L3 affinity, respectively
- **interrupt throughput**: e.g., execute network-facing daemon on same core that processes network interrupts
- **interrupt isolation**: e.g., execute only on half of cores that do not handle device interrupts
- **security isolation**: e.g., avoid micro-architectural timing channels by forcing sensitive and less trusted tasks to run on separate cores

All resulting affinities naturally exhibit structure.

*They are **not completely arbitrary!***

Natural Affinity Structure

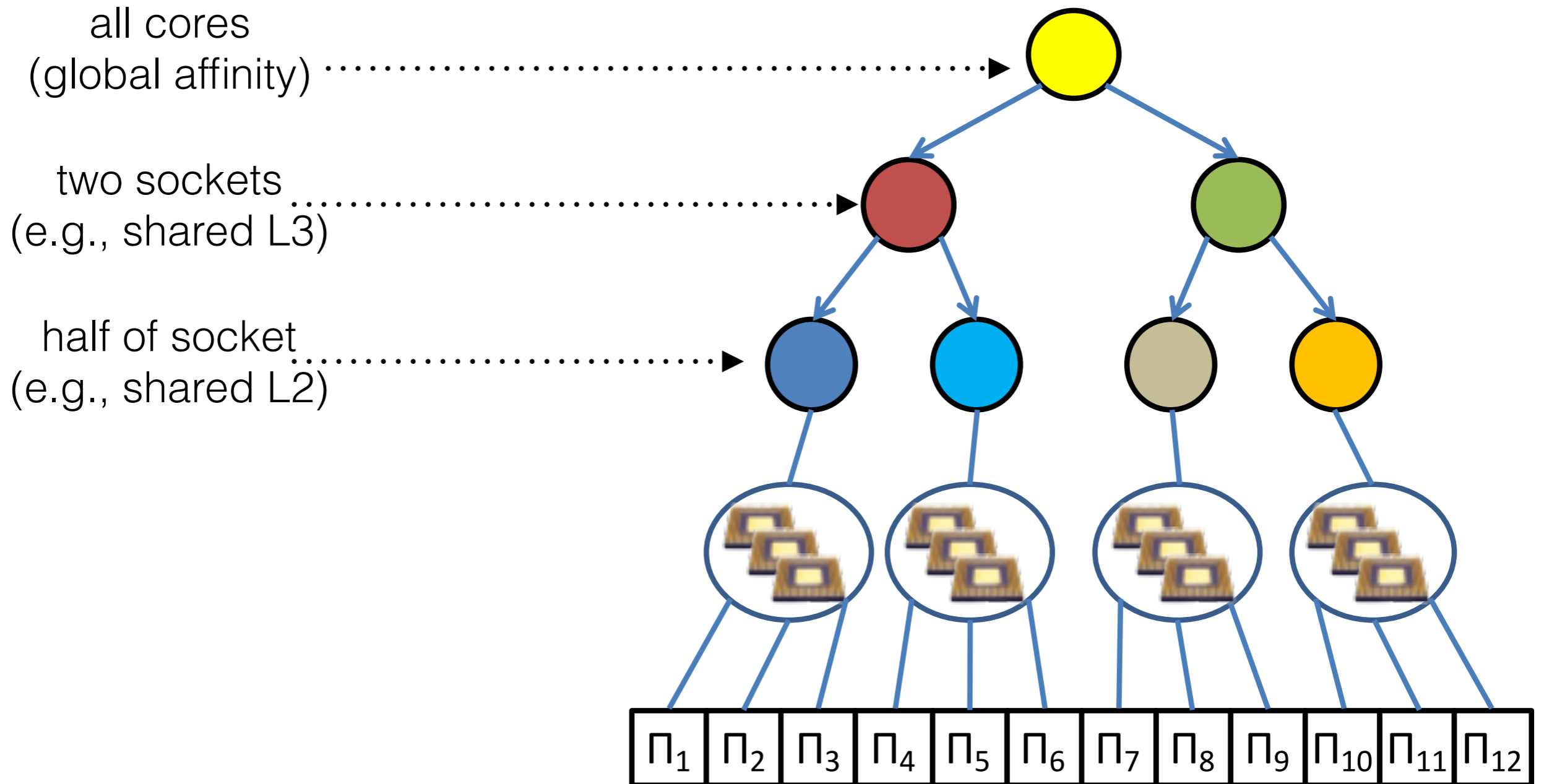
- **Goal: isolation**
 - system sliced into differently sized "compartments"
 - affinities do not overlap (**complete exclusion**)
- **Goal: cache affinity**
 - affinities reflect **memory hierarchy**
 - smaller affinities part of larger affinities (**full inclusion**)
- **Goal: sequencing of tasks (partial partitioning)**
 - singleton affinities
- **Goal: average-case response-time improvements**
 - global (or at least very large) affinities

Hierarchical (or Laminar) Processor Affinities (HPA)

- **Laminar family** of affinity sets (*tree-like structure*)
- For any two jobs i and j , either:

$$\alpha_i \subseteq \alpha_j \quad \text{or} \quad \alpha_j \subseteq \alpha_i \quad \text{or} \quad \alpha_j \cap \alpha_i = \emptyset$$

Example HPA Inclusion Tree



Overview of Results

	Strong APA (<i>Gujarati et al., 2014</i>)	Strong HPA (<i>this paper</i>)
Job arrival cost	$O(m^2)$	$O(m)$
Job departure cost	$O(nm)$	$O(\log n + m^2)$
Speed-up bound	—	2.415 (bi-level + EDF) 3.562 (clustered + EDF)
Implemented in OS?	—	LITMUS ^{RT}
Schedulability test	sufficient	— [<i>prior APA test applies</i>]

n ...number of tasks

m ...number of cores

An Efficient **Strong** **HPA** Scheduler

Insight: Separate Job *Selection* from Job *Placement*

- **Job selection** (or admission): determine the set of jobs that should receive processor service
 - at most m , but subject to affinity constraints.
- **Job placement**: map set of selected jobs to processors, while respecting
 - all affinity constraints and
 - the strong APA invariant.

Algorithms in Paper

- Algorithms 1 & 2: **conceptual** scheduling algorithm
→ proof of *strong APA invariant*, but bad complexity
- Algorithms 3–5: **runtime** scheduling algorithm
→ same schedule, but better complexity
- Algorithm 6: **locality-aware** assignment algorithm
→ avoids some migrations, but worse complexity
→ better suited for kernel-level implementation

Algorithms in Paper

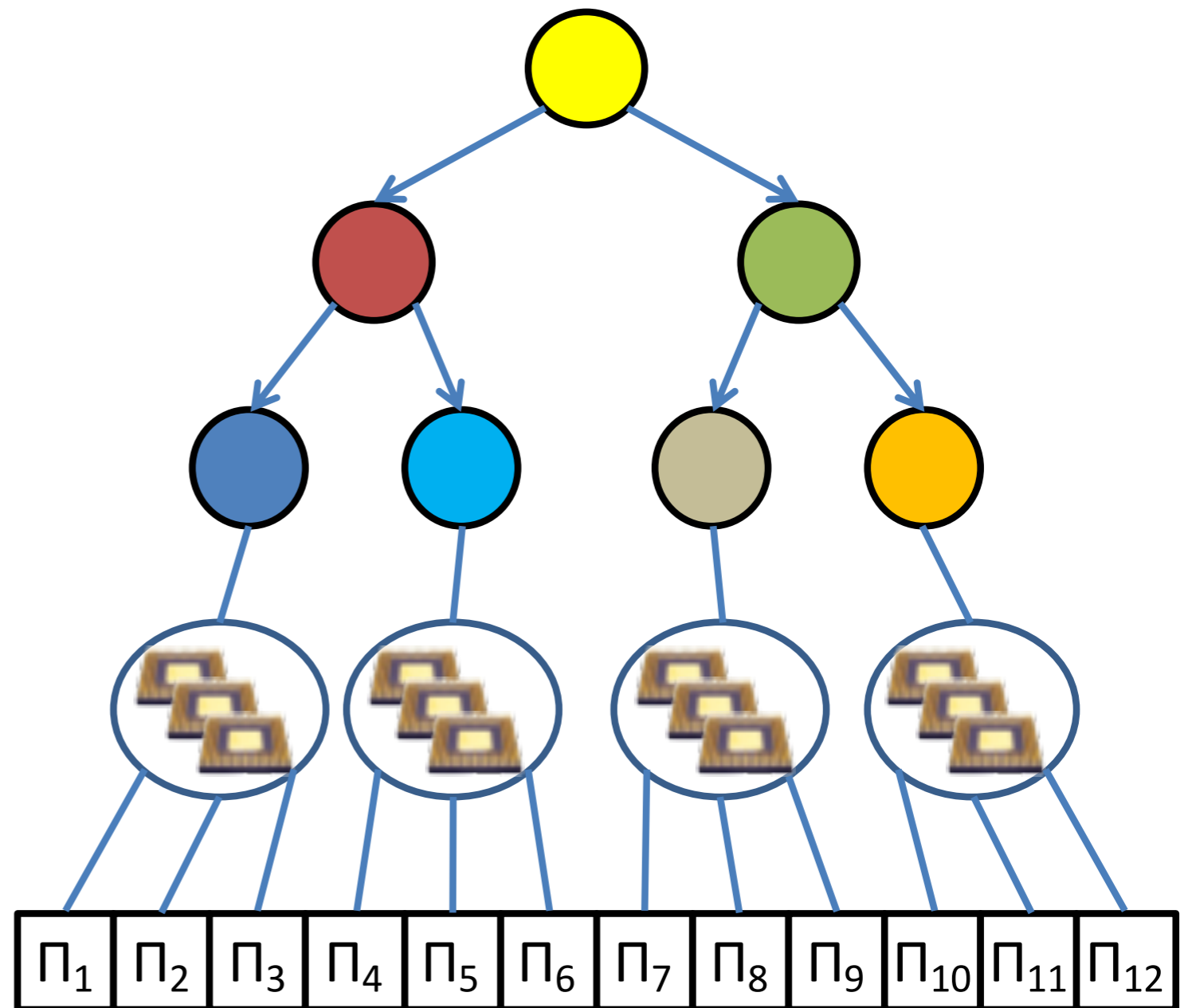
- Algorithms 1 & 2: **conceptual** scheduling algorithm
→ proof of *strong APA invariant*, but bad complexity

- Algorithms 3–5: **runtime** scheduling algorithm
→ same schedule, but better complexity [this talk]

- Algorithm 6: **locality-aware** assignment algorithm
→ avoids some migrations, but worse complexity
→ better suited for kernel-level implementation

Insight: Maintain State for each Distinct Affinity Set

- **don't** have per-processor run-queues (Linux, etc.)
- **don't** have just a single run queue
- **instead**, associate state with each distinct affinity (*affinity tree node*)

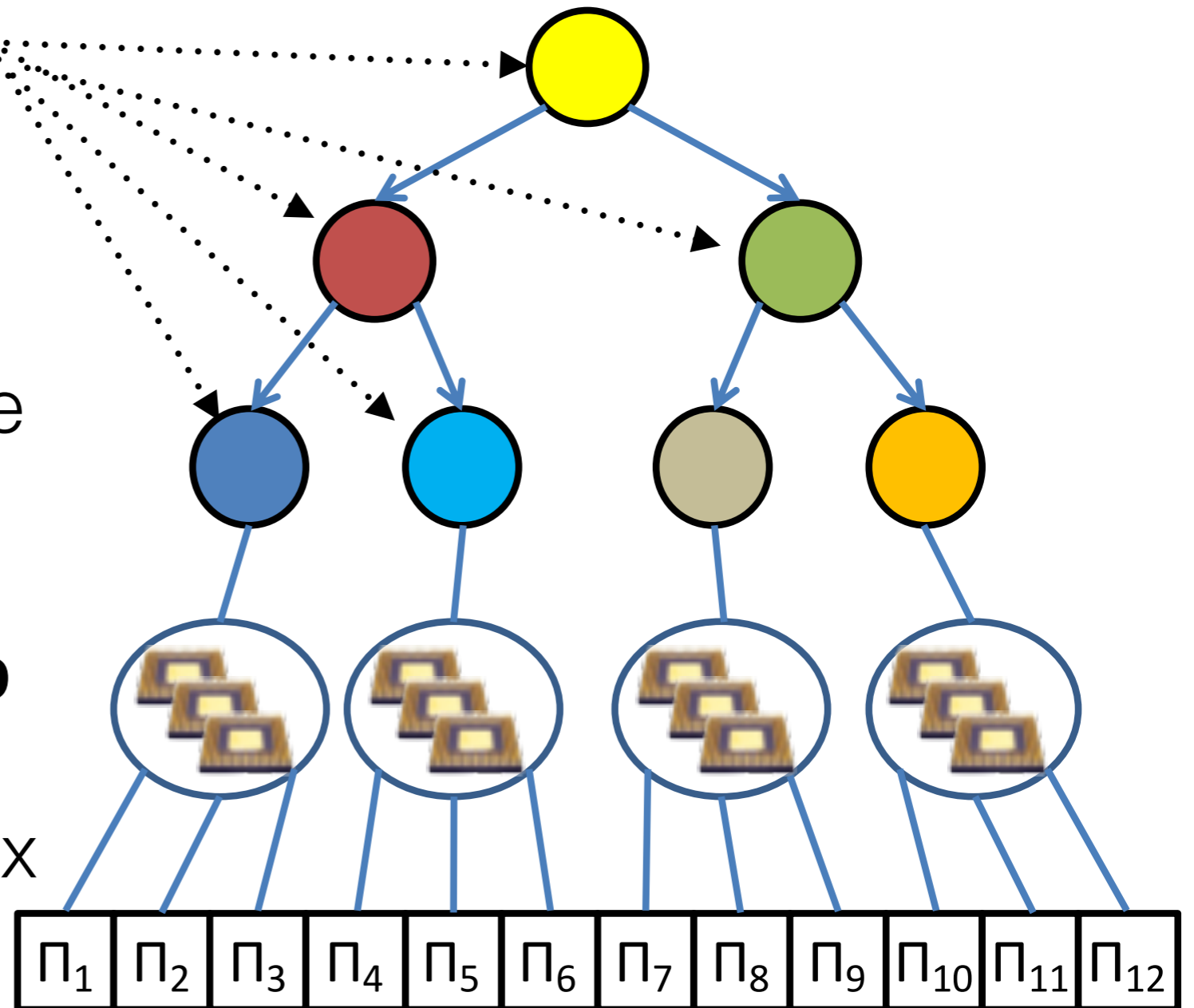


Data Structures

For each distinct affinity

- **doubly linked list** of *scheduled* jobs
 - $O(1)$ Insert, Remove
 - $O(n)$ FindMax

- **strict Fibonacci heap** of *backlogged* jobs
 - $O(1)$ Insert, FindMax
 - $O(\log n)$ Remove

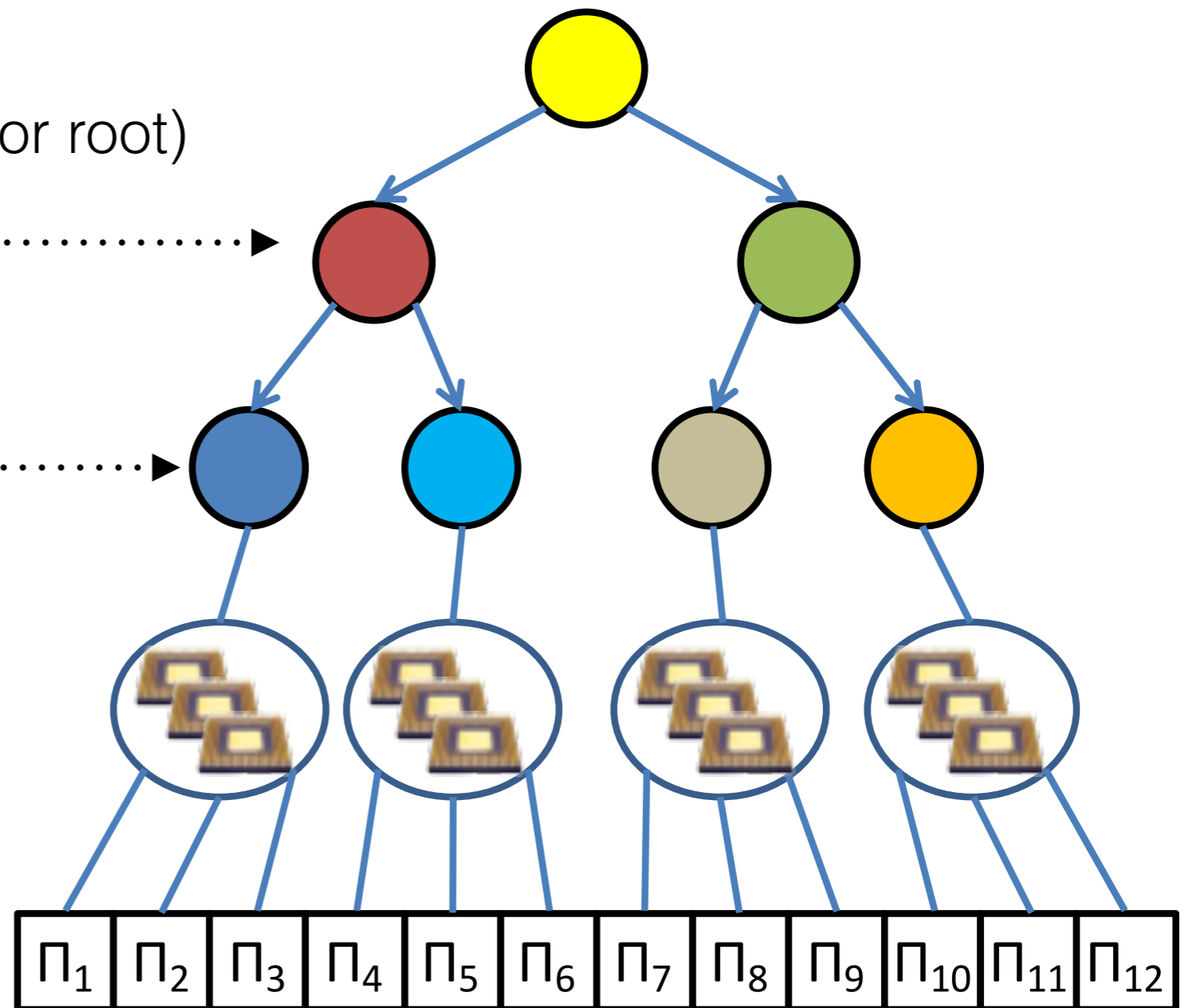


Job Arrival Step 1: Find Beta

first “full” affinity on path to root (or root)

β

α_i
affinity of arriving job

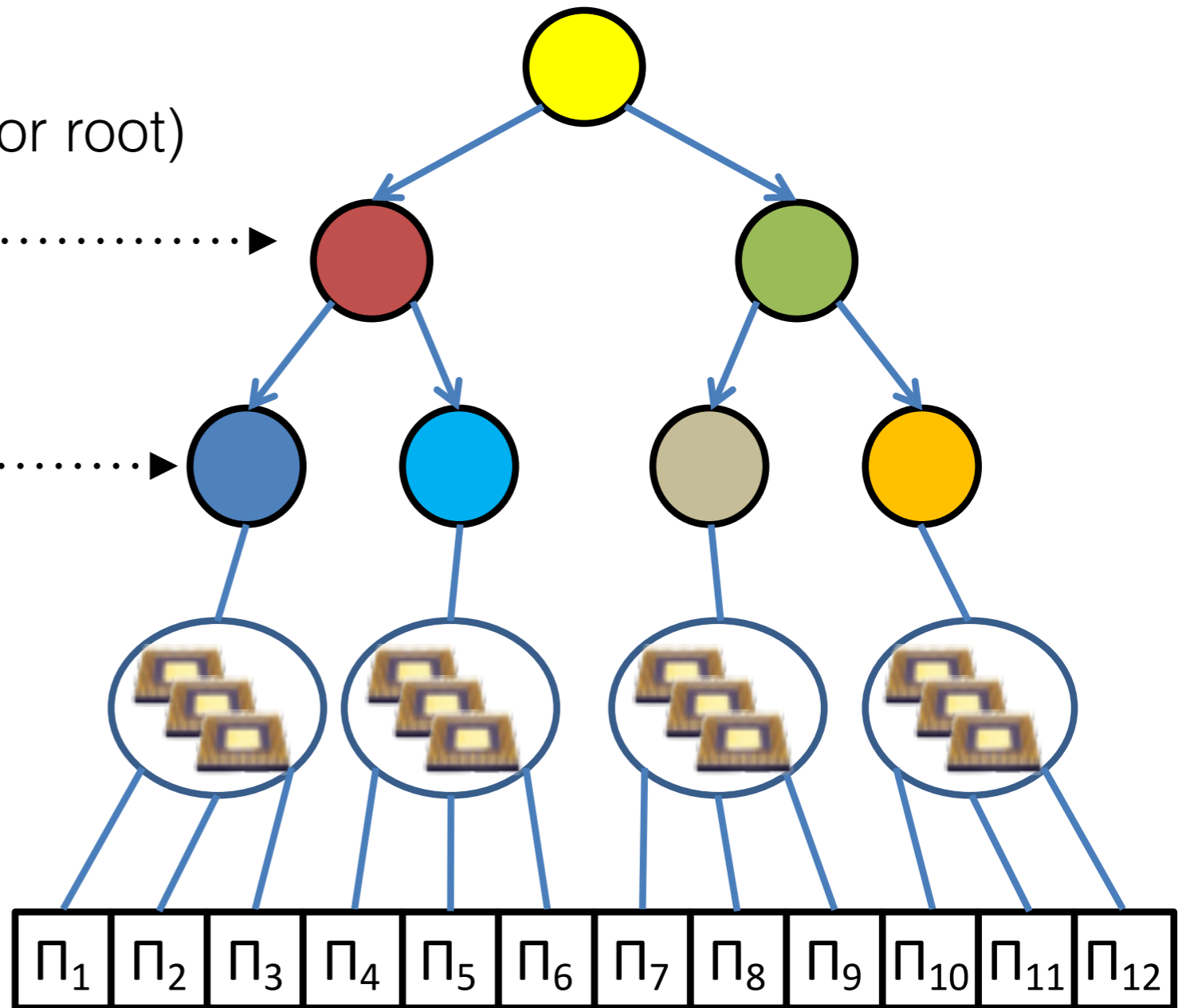


full: #scheduled jobs (list) = #processors in affinity

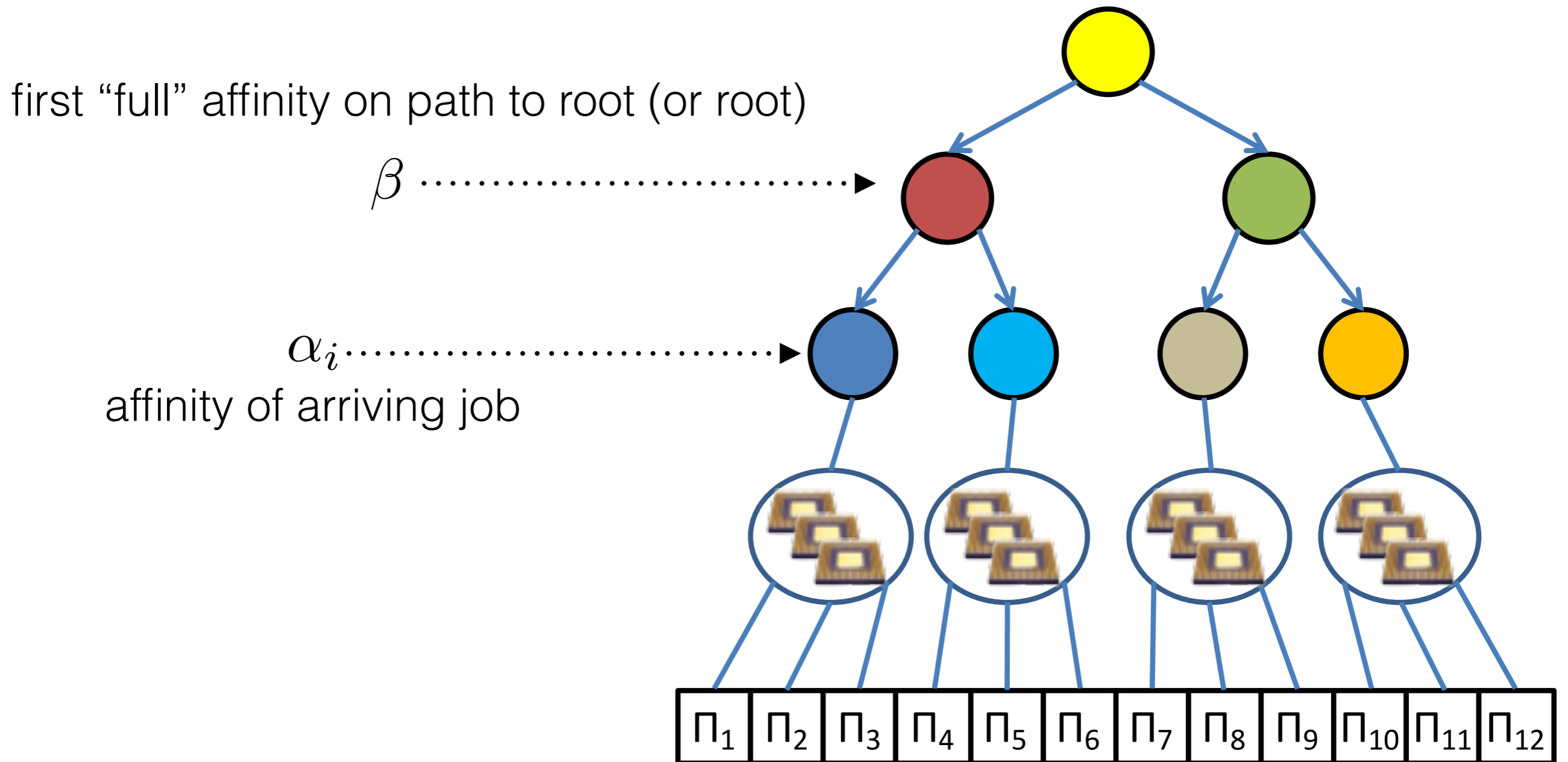
first "full" affinity on path to root (or root)

β

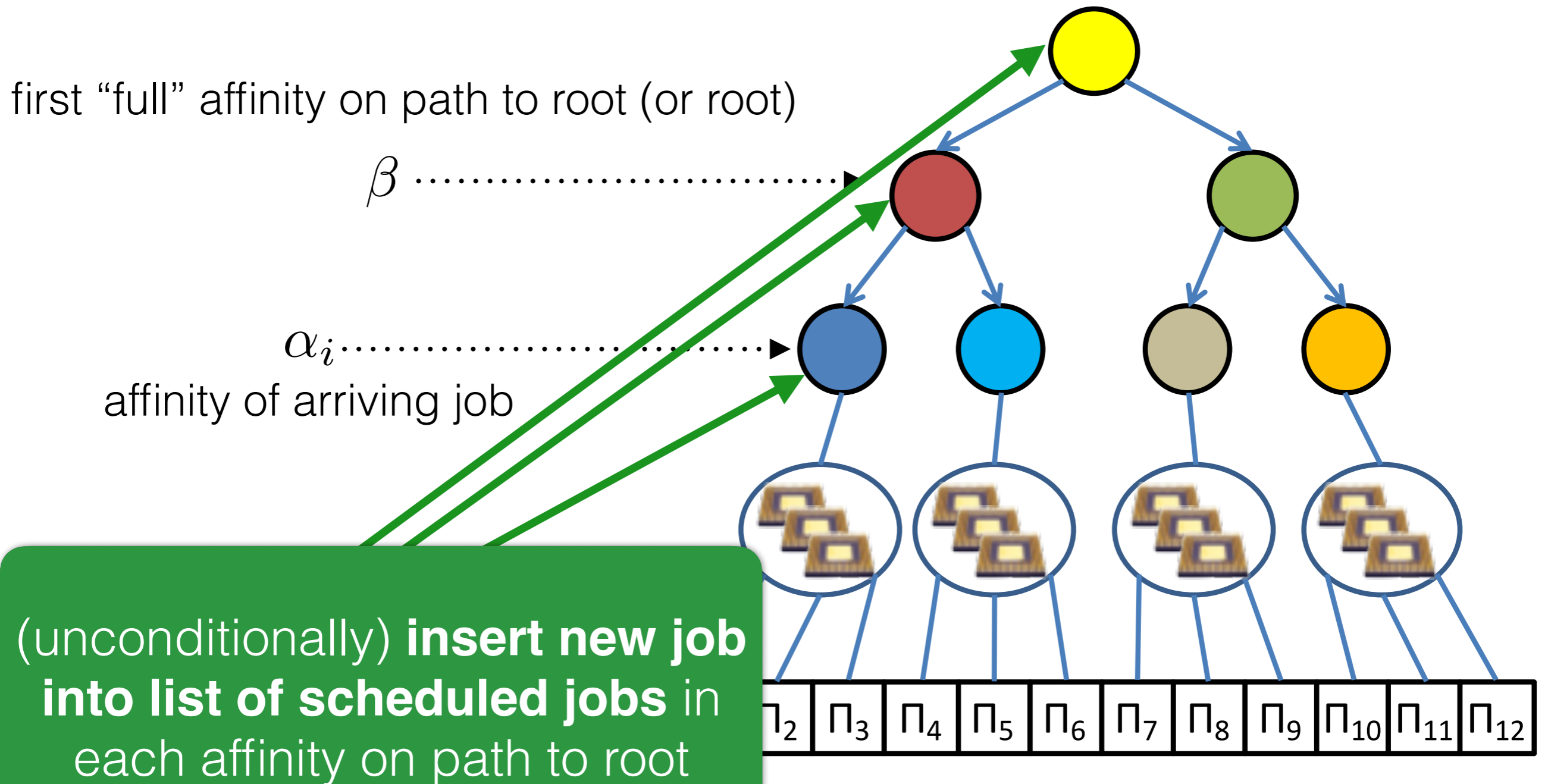
α_i
affinity of arriving job



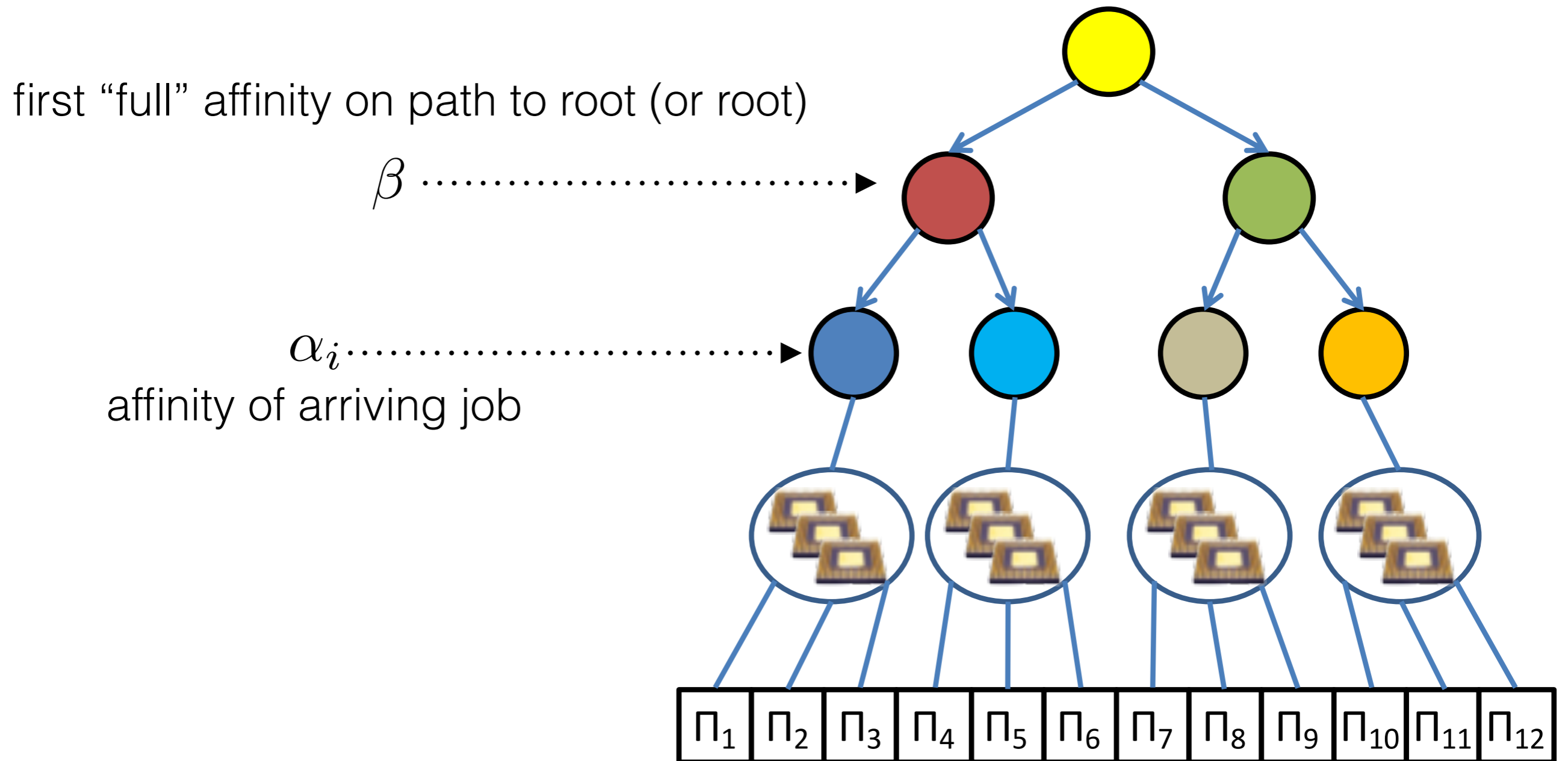
Job Arrival Step 2: Walk Up the Tree and Insert into Lists



Job Arrival Step 2: Walk Up the Tree and Insert into Lists



Job Arrival Step 3: Find Lowest-Priority Job in Beta Affinity's List



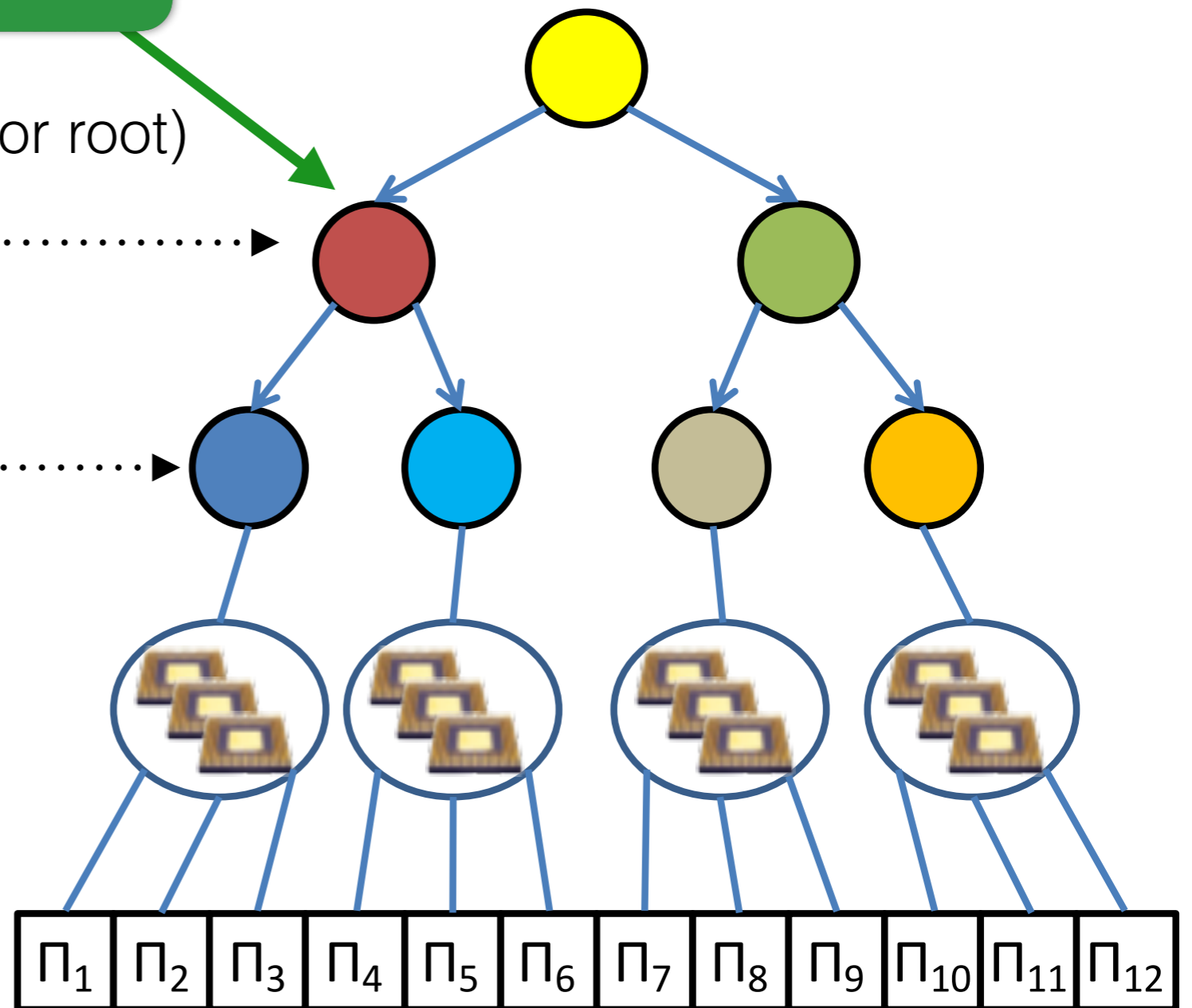
Search list of “scheduled” jobs
to **find lowest-priority job**

3: Find Lowest-eta Affinity's List

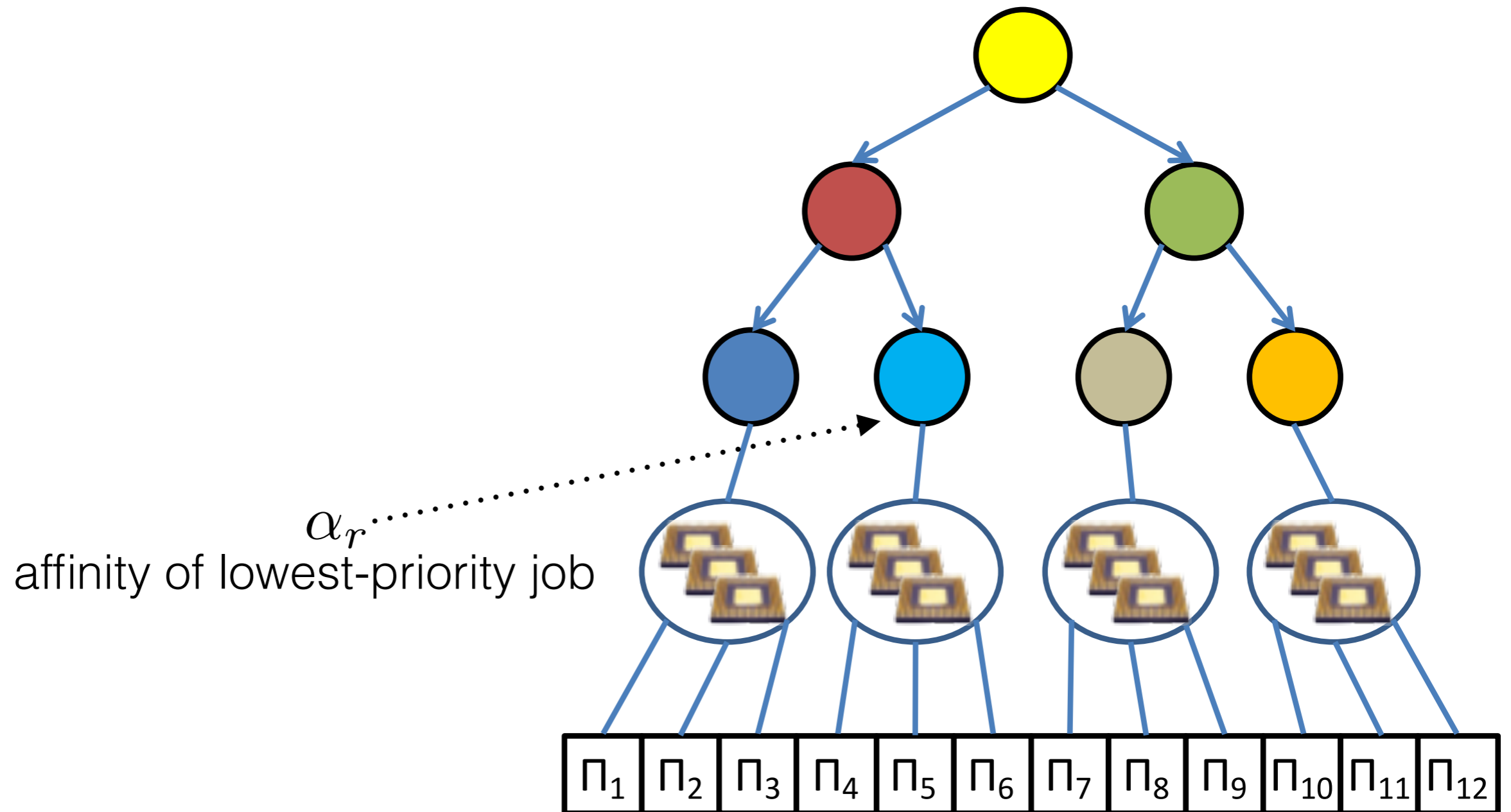
first “full” affinity on path to root (or root)

β

α_i
affinity of arriving job

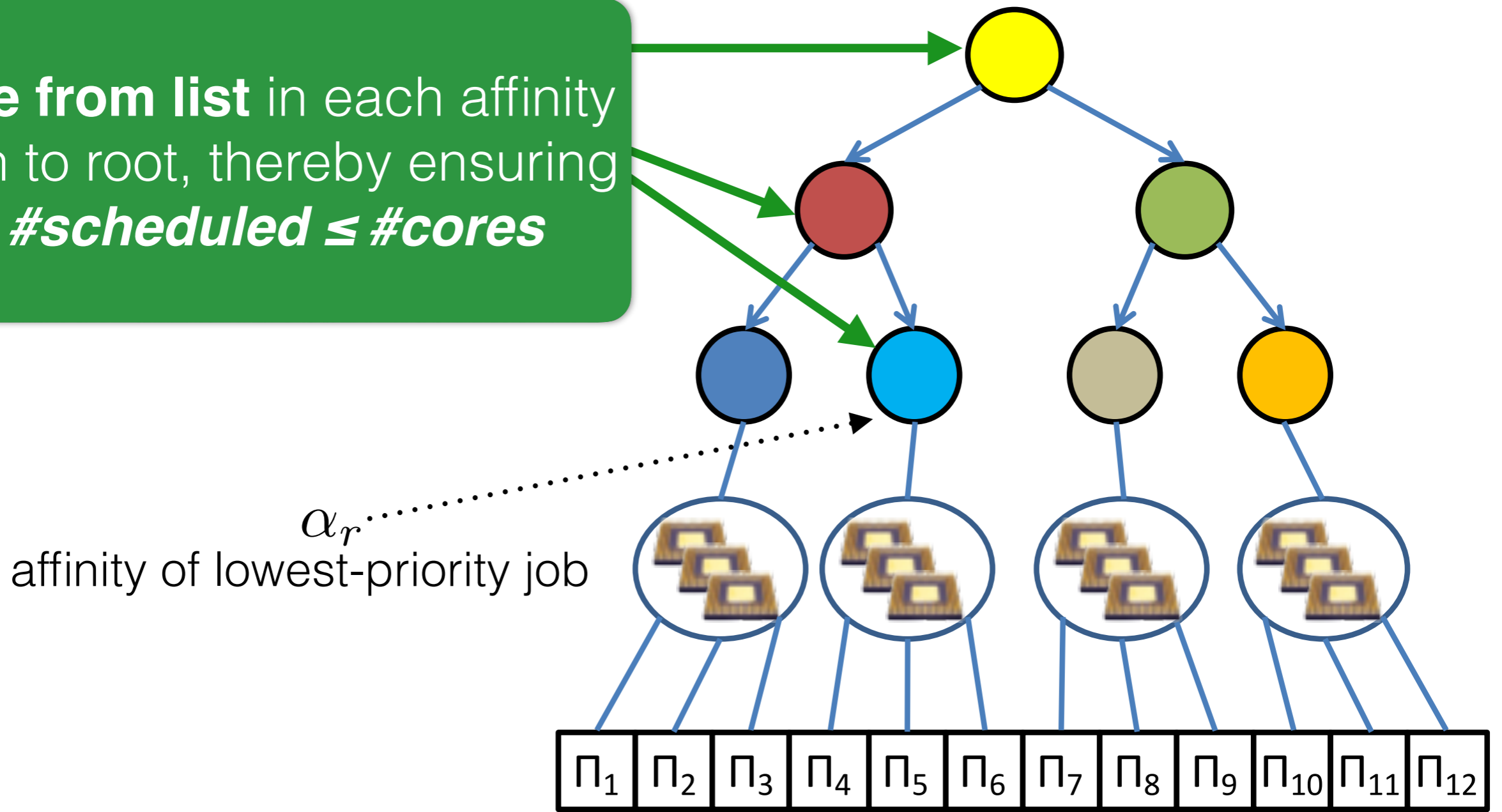


Job Arrival Step 4: Clean Up Queues and Add to Backlogged



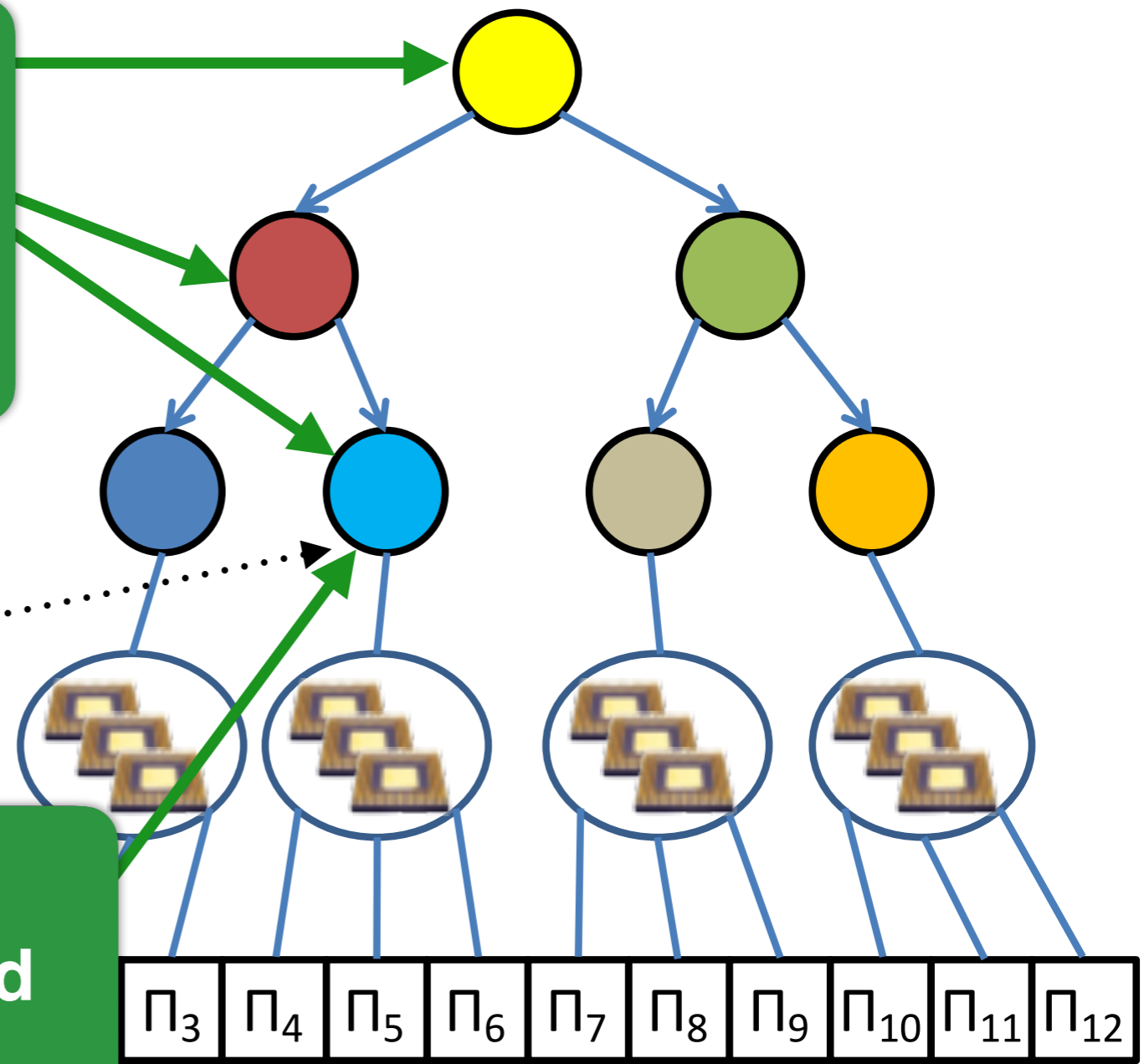
Job Arrival Step 4: Clean Up Queues and Add to Backlogged

remove from list in each affinity on path to root, thereby ensuring that $\#scheduled \leq \#cores$



Job Arrival Step 4: Clean Up Queues and Add to Backlogged

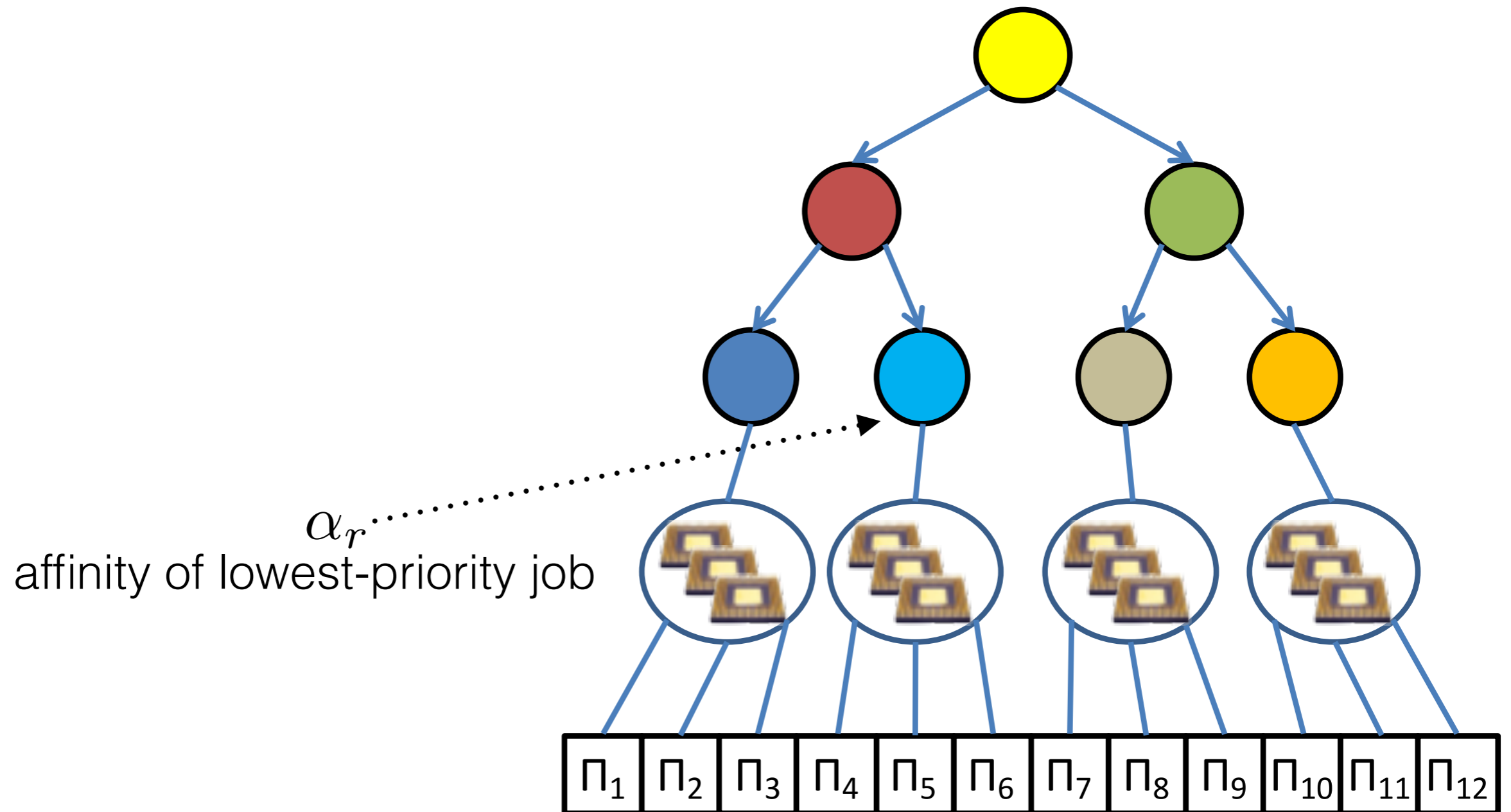
remove from list in each affinity on path to root, thereby ensuring that $\#scheduled \leq \#cores$



α_r
affinity of lowest-priority job

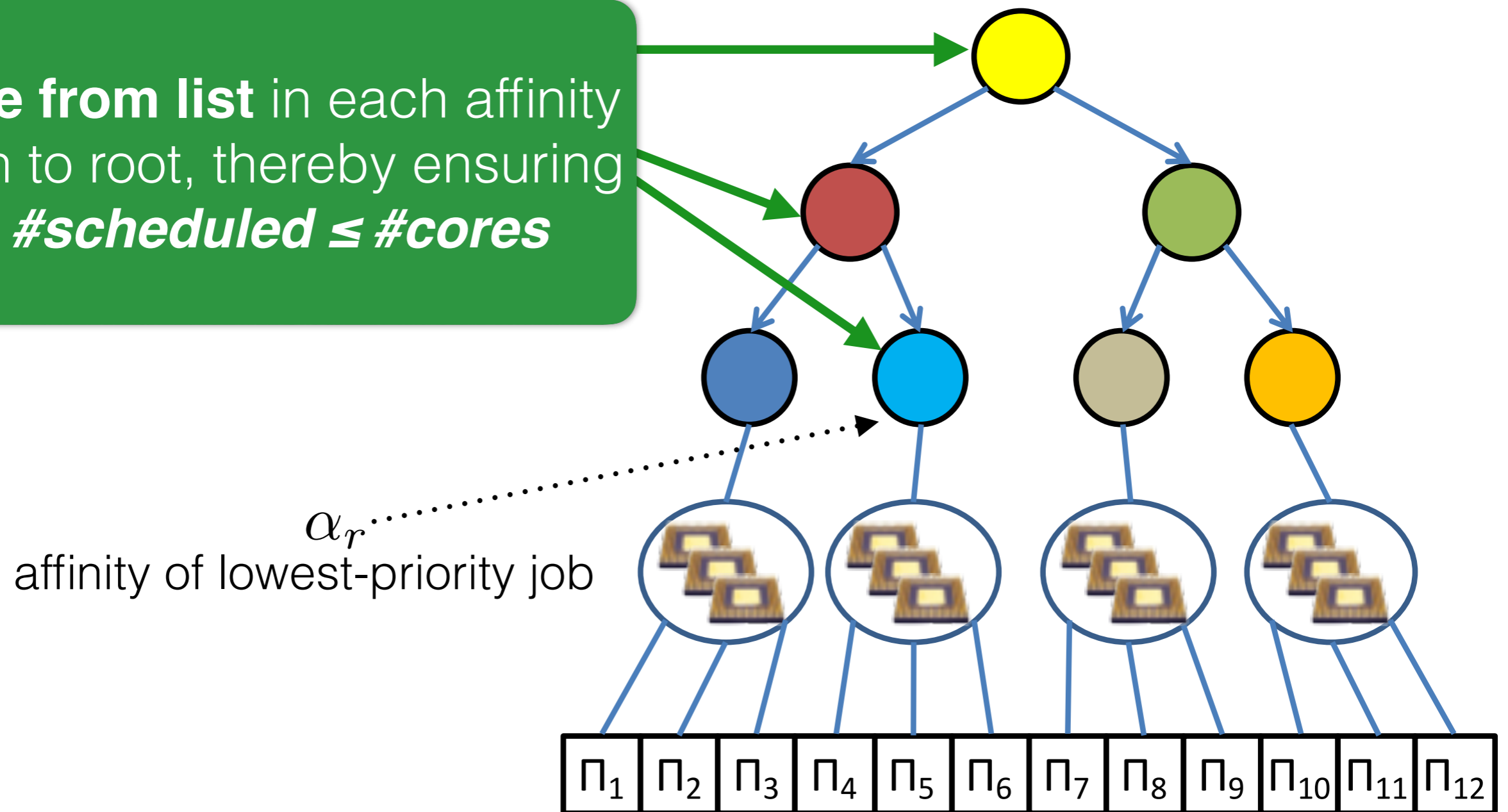
add to heap of backlogged jobs (only in own affinity)

Job Arrival Step 4: Clean Up Lists along Path to Root

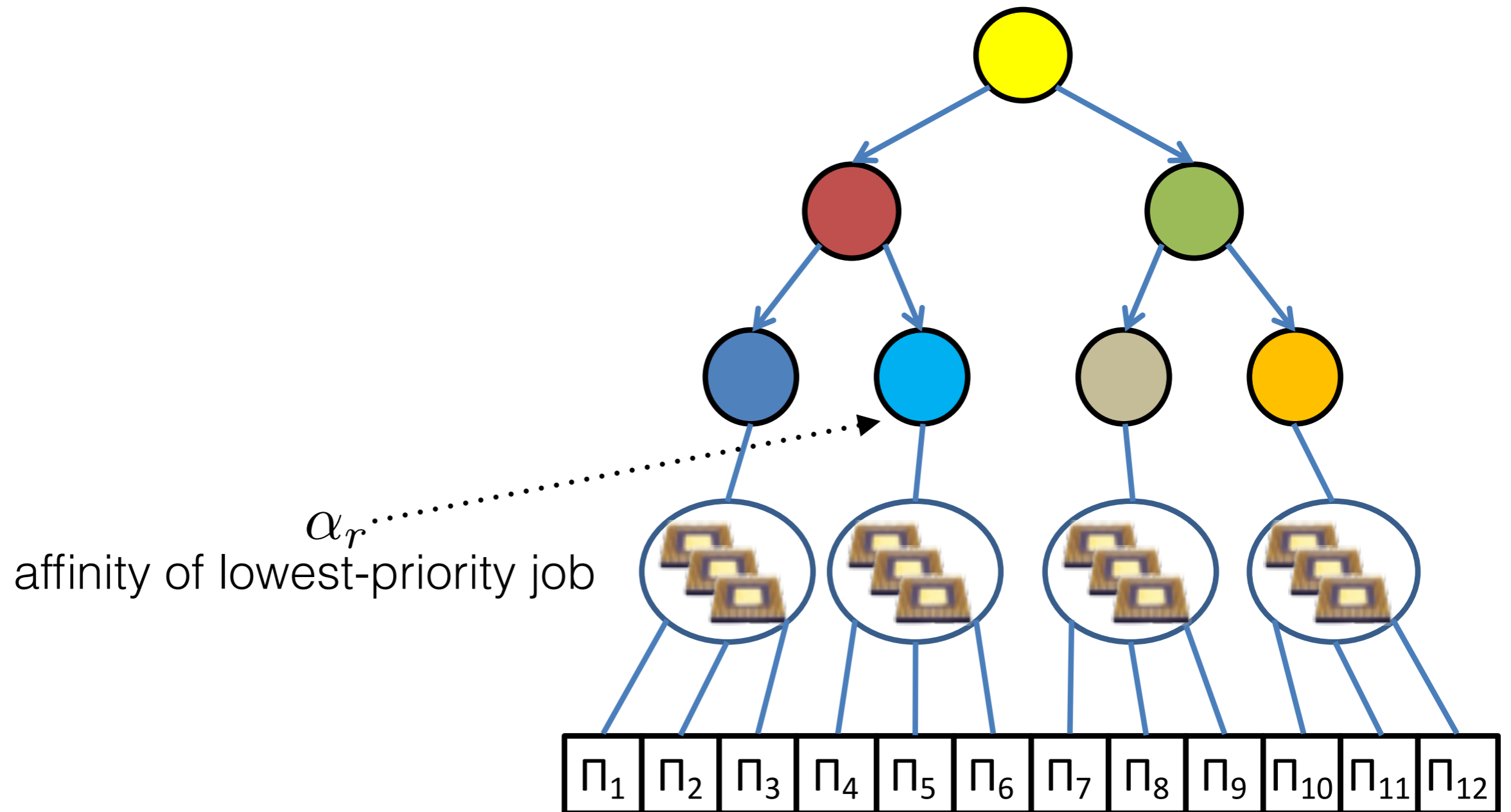


Job Arrival Step 4: Clean Up Lists along Path to Root

remove from list in each affinity on path to root, thereby ensuring that $\#scheduled \leq \#cores$

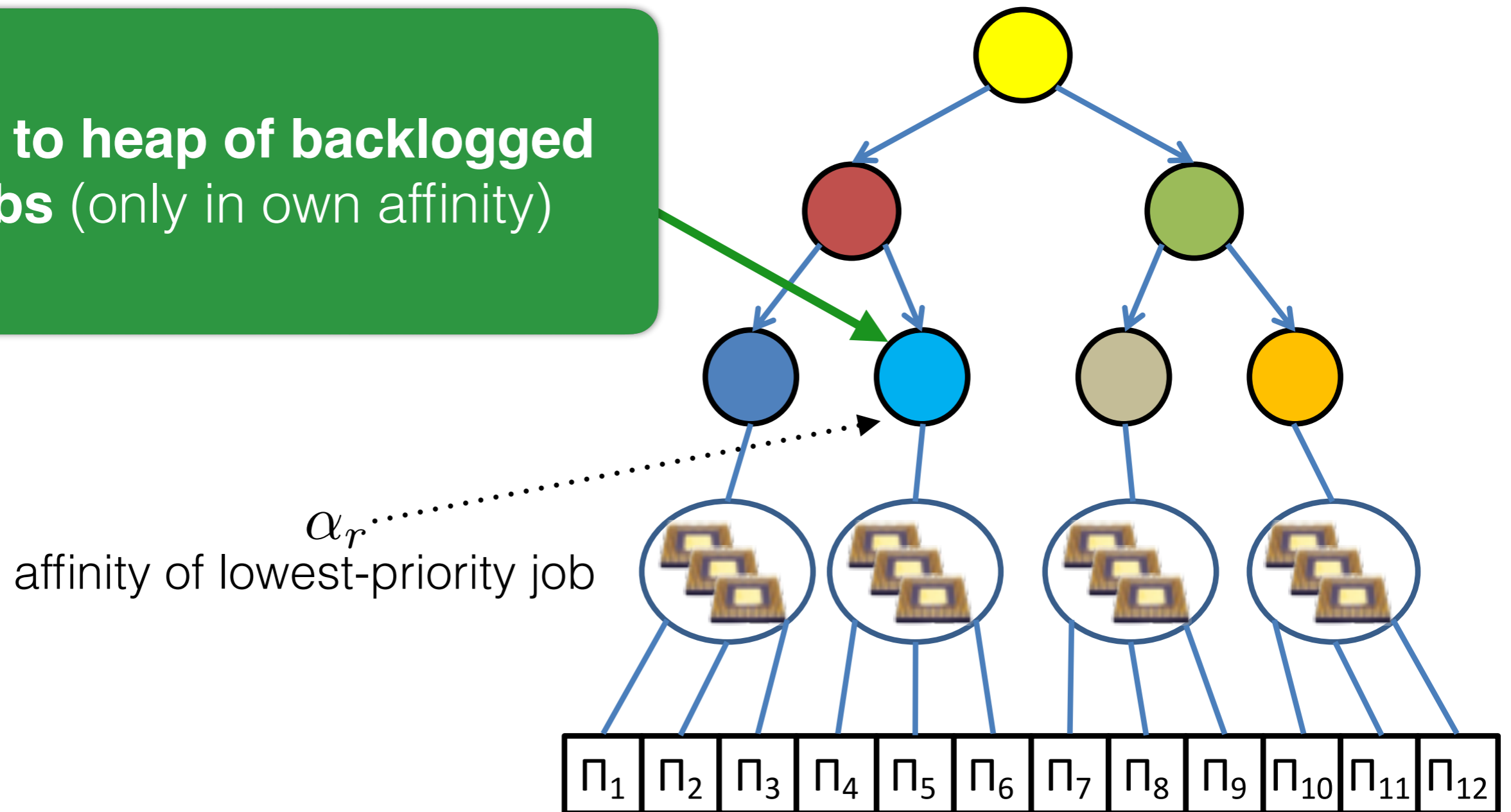


Job Arrival Step 5: Add to Heap of Backlogged Jobs



Job Arrival Step 5: Add to Heap of Backlogged Jobs

add to heap of backlogged jobs (only in own affinity)



Complexity of Job Selection
upon Arrival: $O(m)$

m ...number of cores

Complexity of Job Selection upon Arrival: $O(m)$

1. Find beta affinity, first “full” affinity: $O(\text{height of tree}) = O(m)$

Complexity of Job Selection upon Arrival: $O(m)$

1. Find beta affinity, first “full” affinity: $O(\text{height of tree}) = O(m)$
2. Walk up the tree and insert new job into doubly-linked lists: $O(\text{height of tree}) = O(m)$

Complexity of Job Selection upon Arrival: $O(m)$

1. Find beta affinity, first “full” affinity: $O(\text{height of tree}) = O(m)$
2. Walk up the tree and insert new job into doubly-linked lists: $O(\text{height of tree}) = O(m)$
3. Find lowest-priority job in beta affinity’s list of scheduled jobs: $O(\text{length of list}) = O(\text{size of affinity}) = O(m)$

Complexity of Job Selection upon Arrival: $O(m)$

1. Find beta affinity, first “full” affinity: $O(\text{height of tree}) = O(m)$
2. Walk up the tree and insert new job into doubly-linked lists: $O(\text{height of tree}) = O(m)$
3. Find lowest-priority job in beta affinity’s list of scheduled jobs: $O(\text{length of list}) = O(\text{size of affinity}) = O(m)$
4. Walk up the tree and remove lowest-priority job from doubly-linked lists: $O(\text{height of tree}) = O(m)$

m ...number of cores

Complexity of Job Selection upon Arrival: $O(m)$

1. Find beta affinity, first “full” affinity: $O(\text{height of tree}) = O(m)$
2. Walk up the tree and insert new job into doubly-linked lists: $O(\text{height of tree}) = O(m)$
3. Find lowest-priority job in beta affinity’s list of scheduled jobs: $O(\text{length of list}) = O(\text{size of affinity}) = O(m)$
4. Walk up the tree and remove lowest-priority job from doubly-linked lists: $O(\text{height of tree}) = O(m)$
5. Add to strict Fibonacci heap of backlogged jobs: $O(1)$

m ...number of cores

Job Arrival Part 2: Placing the
Set of Selected Jobs: $O(m)$

Job Arrival Part 2: Placing the Set of Selected Jobs: $O(m)$

1. Make a list for each **leaf node** in the affinity tree, containing the **free processors** in the affinity: $O(m)$

Job Arrival Part 2: Placing the Set of Selected Jobs: $O(m)$

1. Make a list for each **leaf node** in the affinity tree, containing the **free processors** in the affinity: $O(m)$
2. copy the **list of scheduled jobs** from the **root node**: $O(m)$

Job Arrival Part 2: Placing the Set of Selected Jobs: $O(m)$

1. Make a list for each **leaf node** in the affinity tree, containing the **free processors** in the affinity: $O(m)$
2. copy the **list of scheduled jobs** from the **root node**: $O(m)$
3. **sort** the list of jobs by **increasing affinity level**
(= decreasing distance to root node): $O(m)$ — **counting sort**

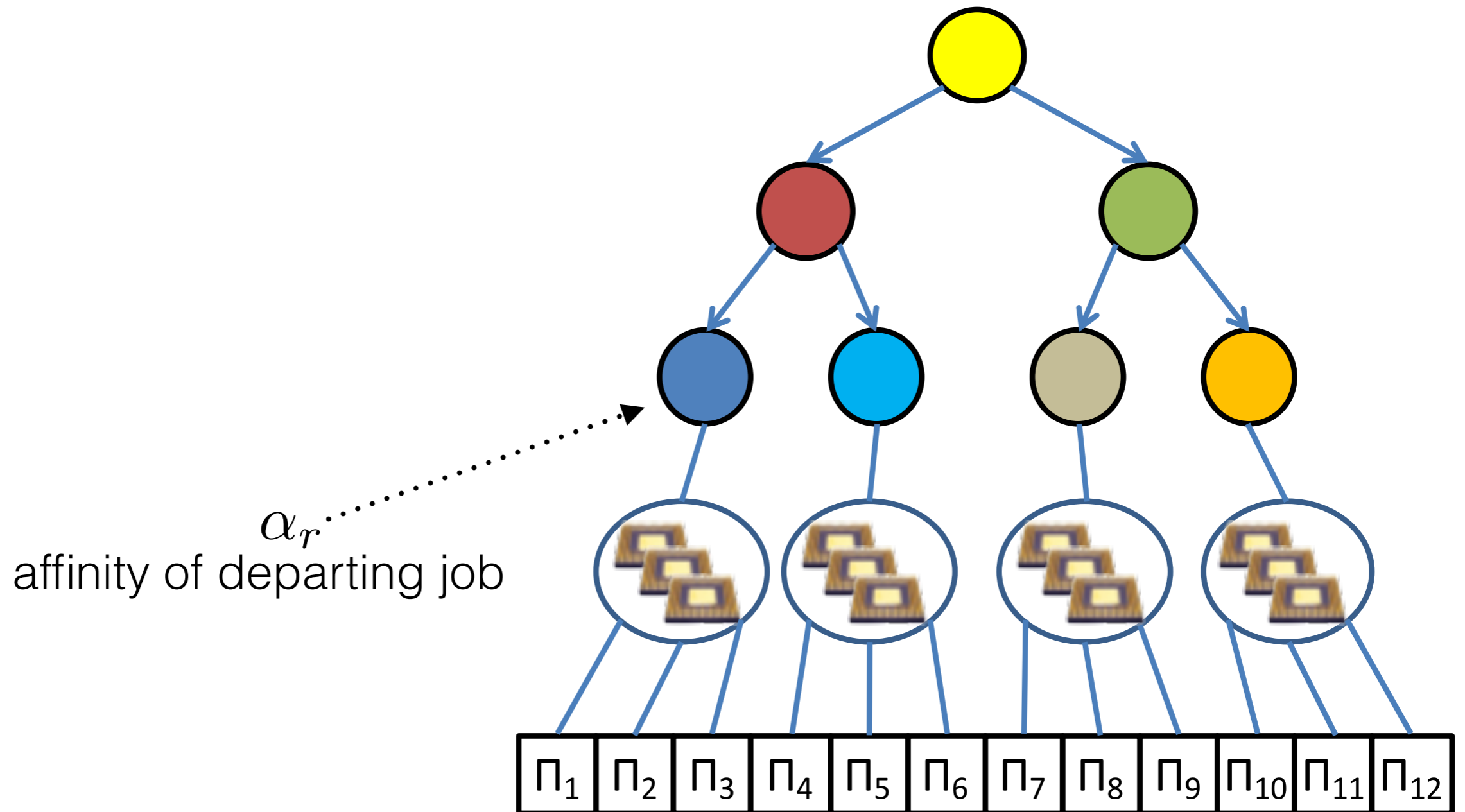
Job Arrival Part 2: Placing the Set of Selected Jobs: $O(m)$

1. Make a list for each **leaf node** in the affinity tree, containing the **free processors** in the affinity: $O(m)$
2. copy the **list of scheduled jobs** from the **root node**: $O(m)$
3. **sort** the list of jobs by **increasing affinity level** (= decreasing distance to root node): $O(m)$ — **counting sort**
4. for each job:
 - assign to **first core** in job affinity's free processor list and remove core from list: $O(m)$
 - when **moving up a level**, concatenate the processor lists of all child nodes and assign to parent node: $O(\text{number of distinct affinities}) = O(m)$

Insight: Reuse Job Arrival Procedure for "Cleanup" After Job Departure

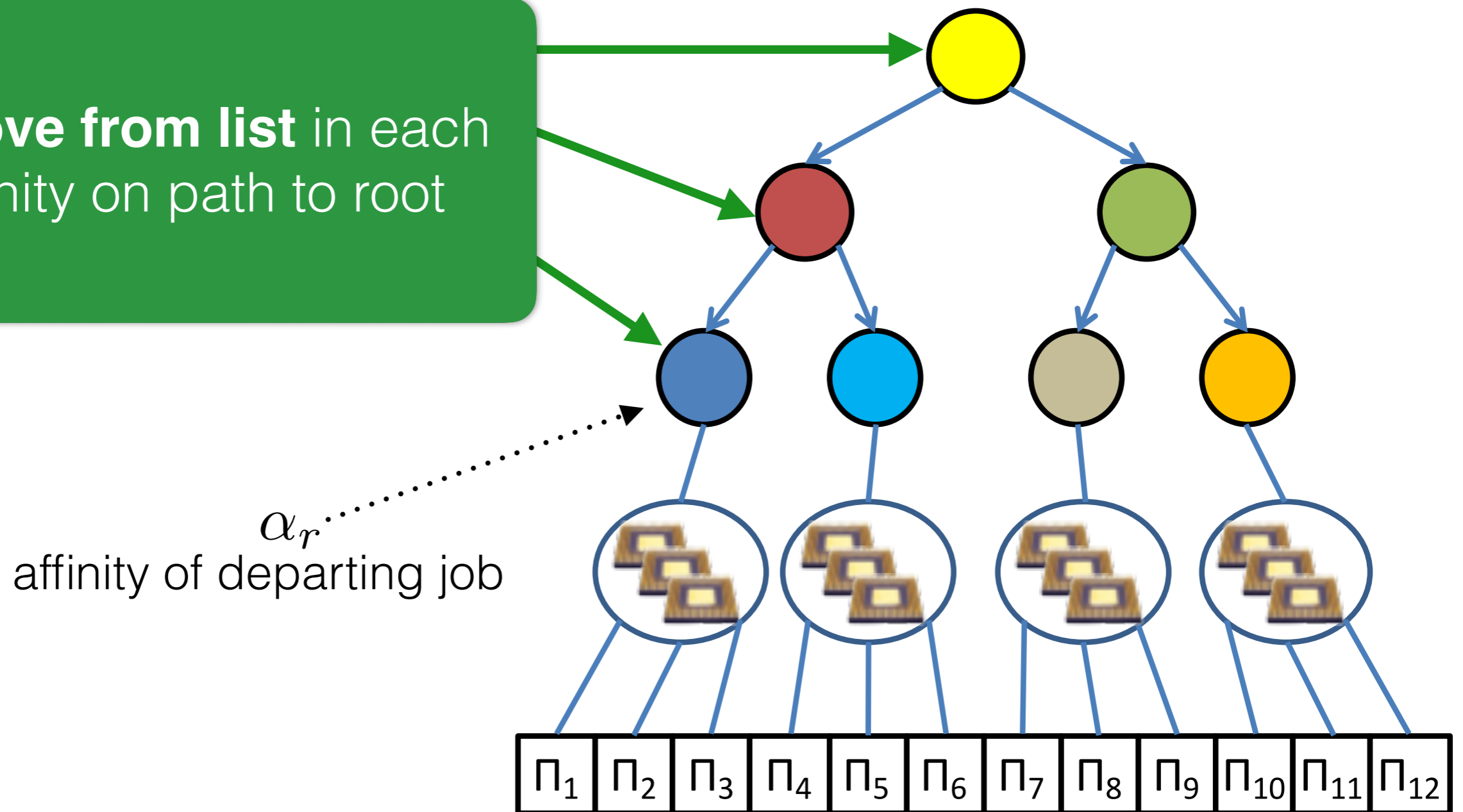
- **Problem:** restoring the strong APA invariant after a job departure is not trivial.
 - If there are backlogged jobs at every affinity node, the **next job to be scheduled** could come from **any** of the affinity nodes in the tree.
- **Solution:** we can reuse the job arrival procedure if we “**simulate**” a job **arrival** of the highest-priority backlogged job for each distinct affinity

Job Departure Step 1: Remove from Lists

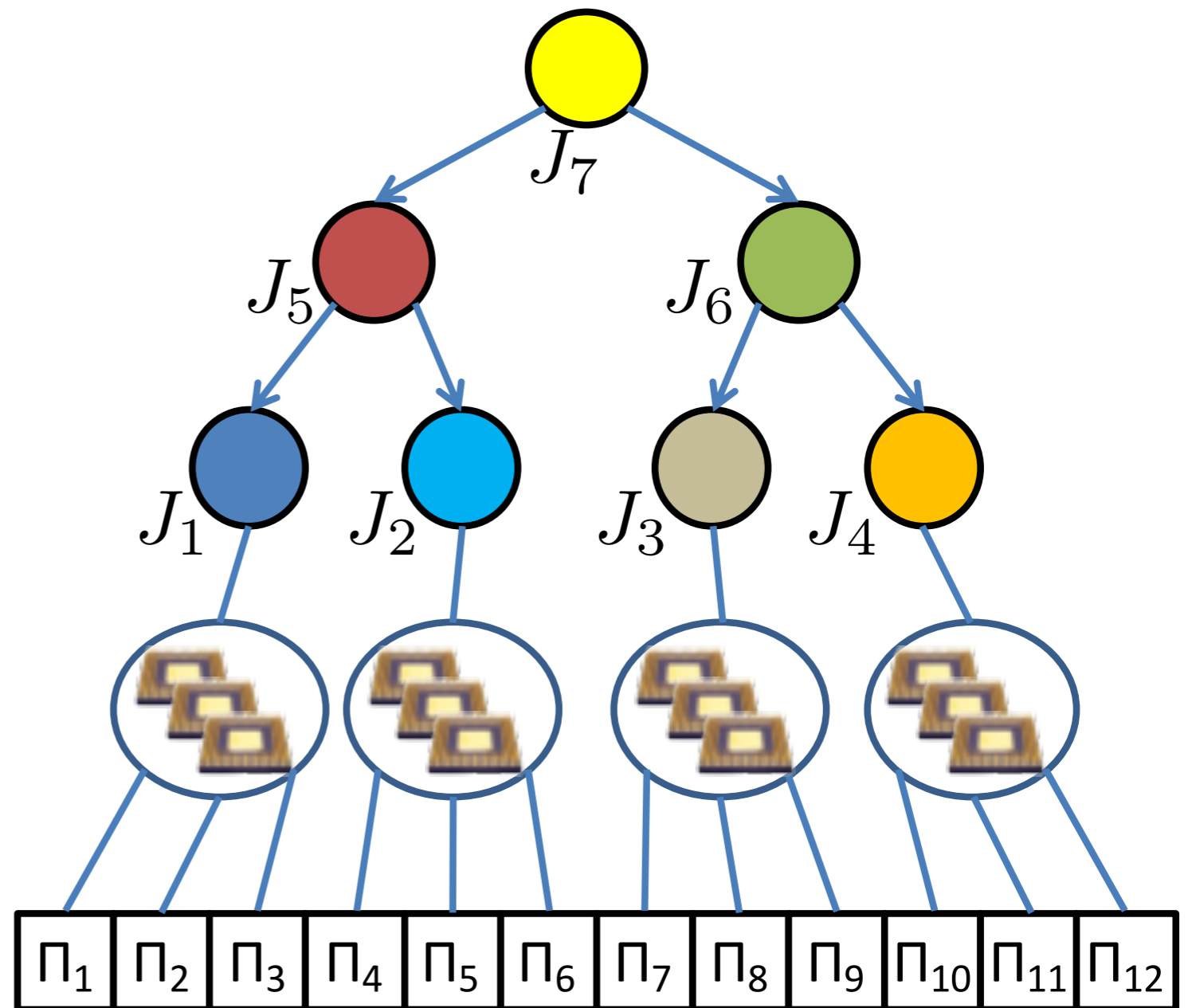


Job Departure Step 1: Remove from Lists

remove from list in each
affinity on path to root

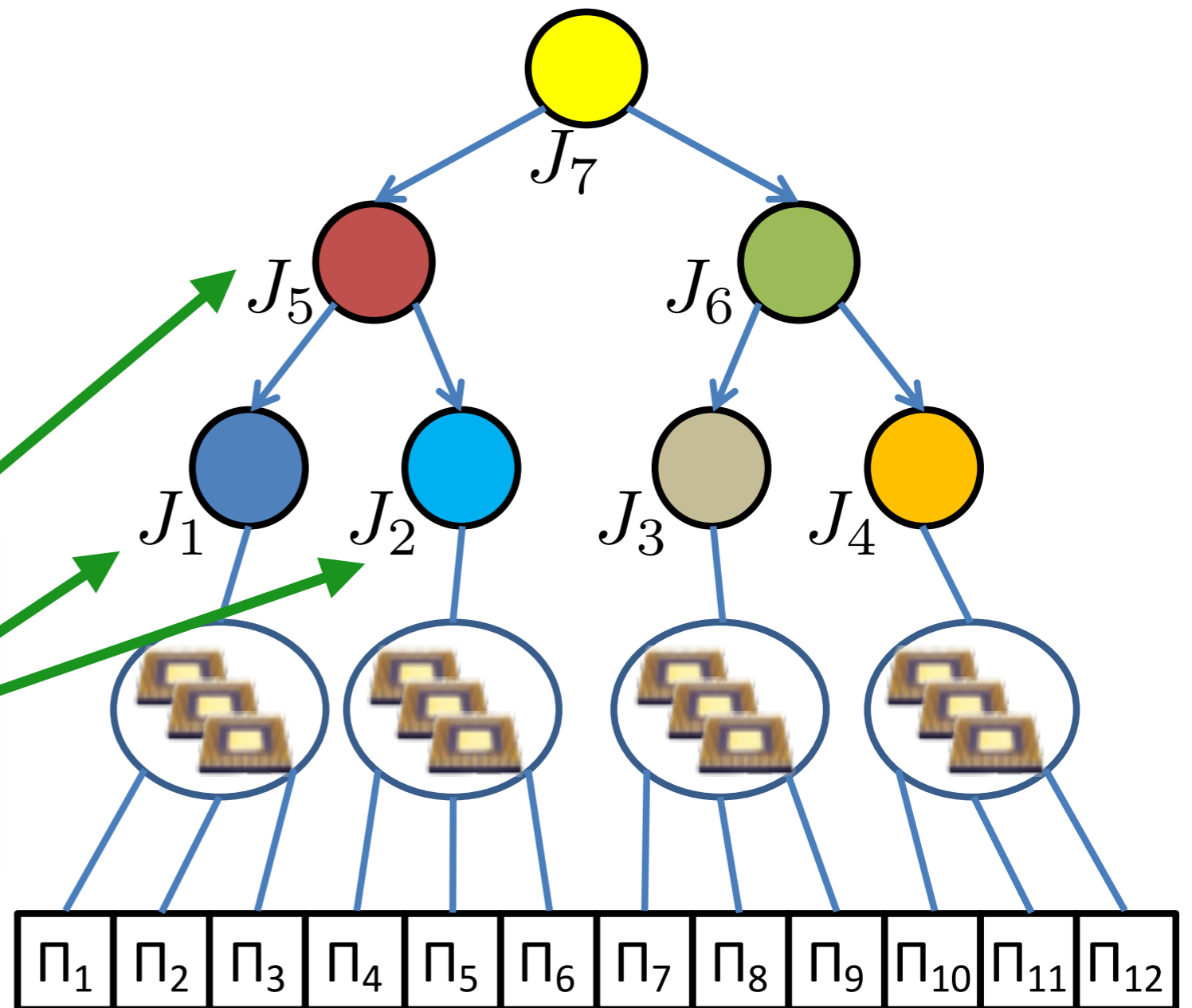


Job Departure Step 2: Find Max in each Affinity



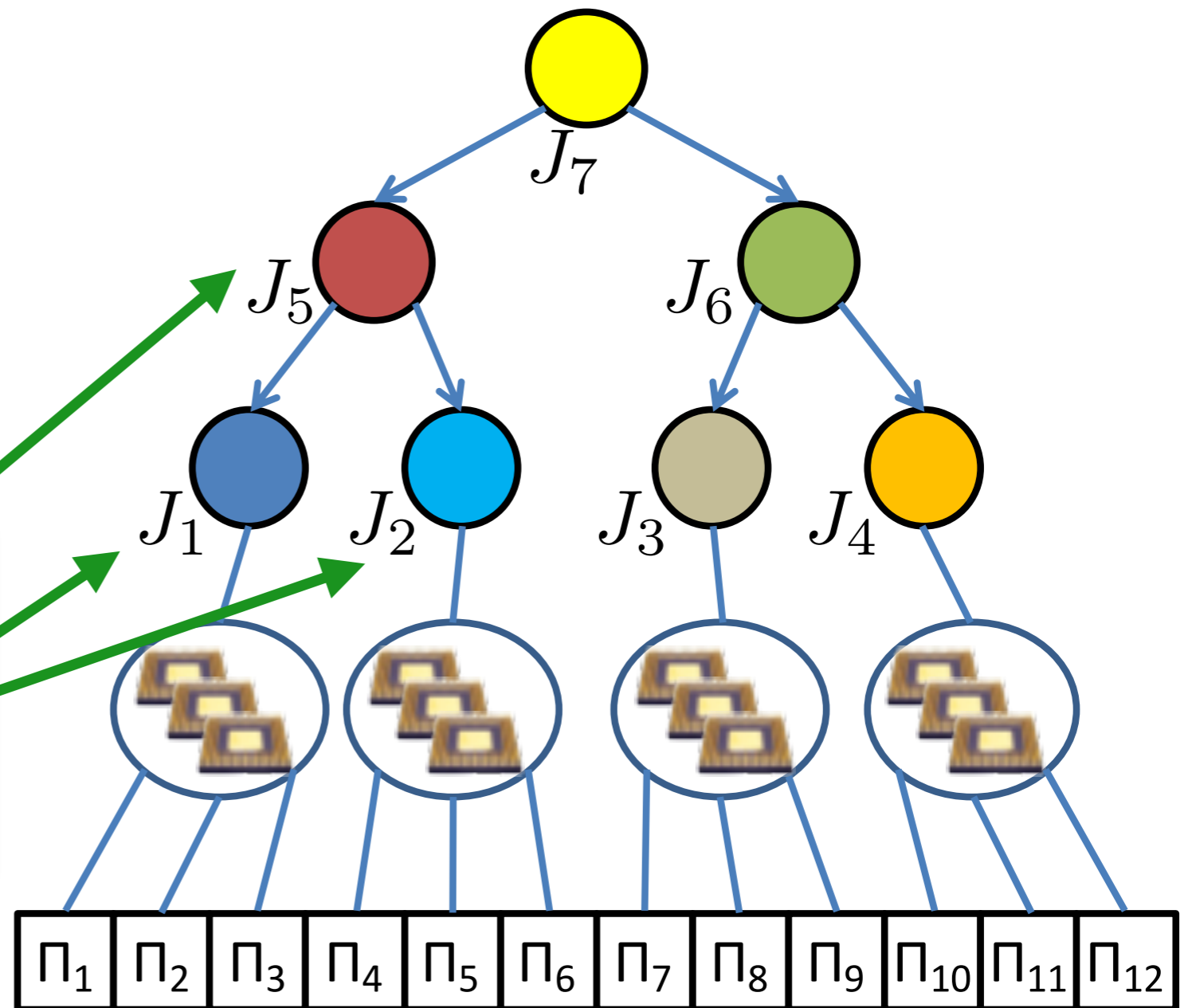
Job Departure Step 2: Find Max in each Affinity

find **highest-priority backlogged job** in each distinct affinity (Fibonacci Heap)



Job Departure Step 3: Simulate Arrivals

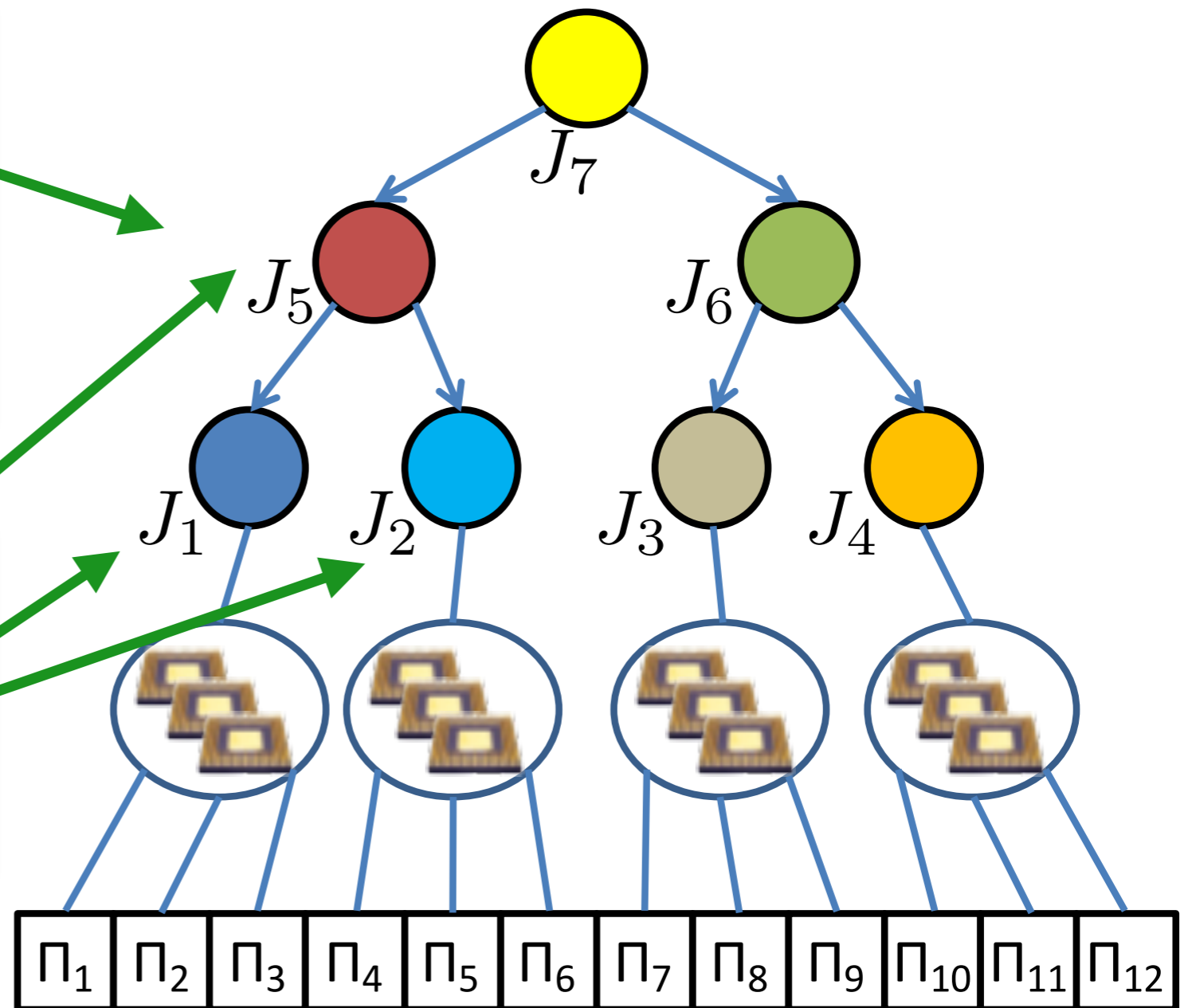
find **highest-priority
backlogged job** in each
distinct affinity
(Fibonacci Heap)



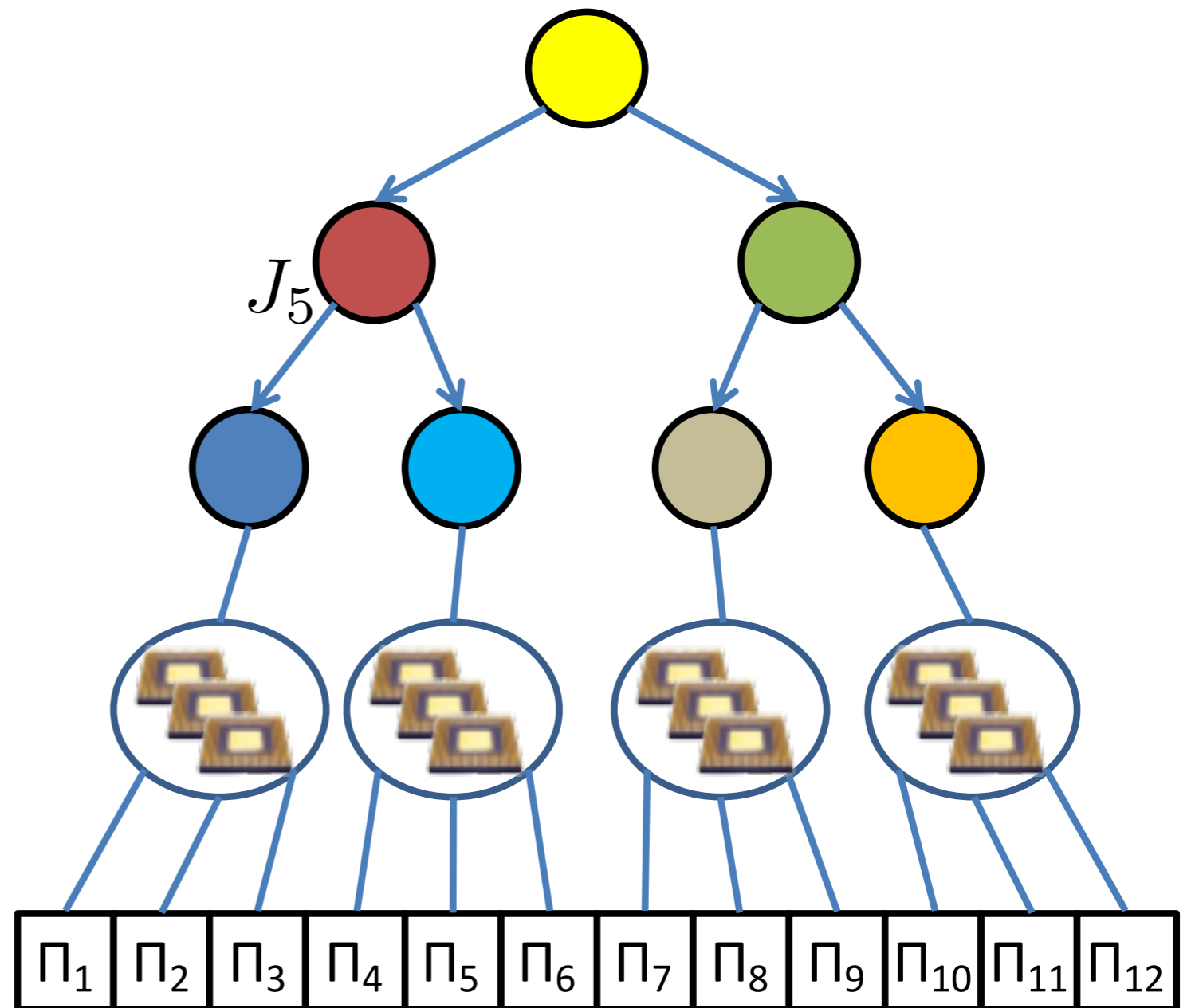
Job Departure Step 3: Simulate Arrivals

run **arrival procedure** for each such job (in any order)
[but don't modify backlogged heap]

find **highest-priority backlogged job** in each distinct affinity
(Fibonacci Heap)

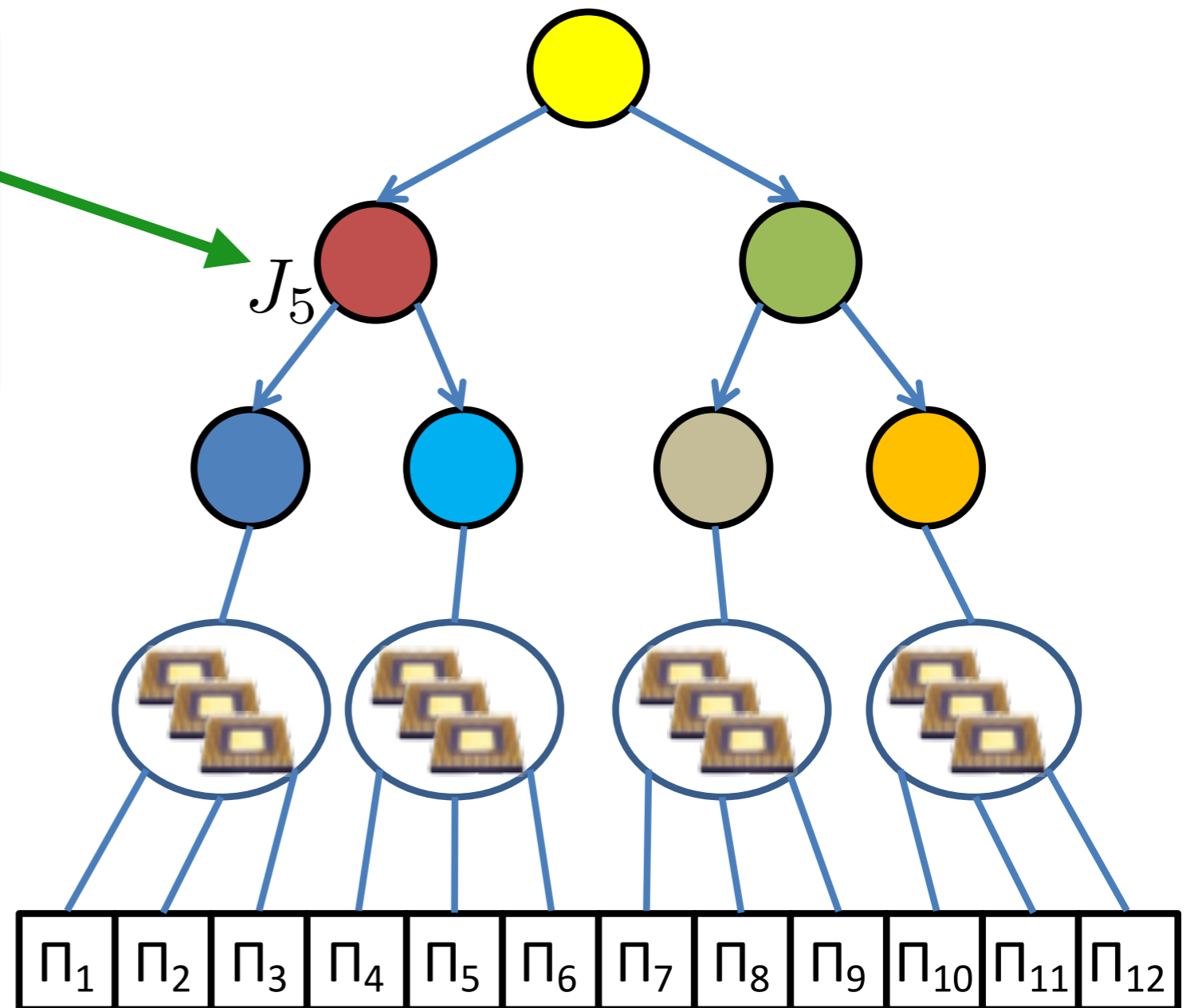


Job Departure Step 4: Remove from Backlogged Heap



Job Departure Step 4: Remove from Backlogged Heap

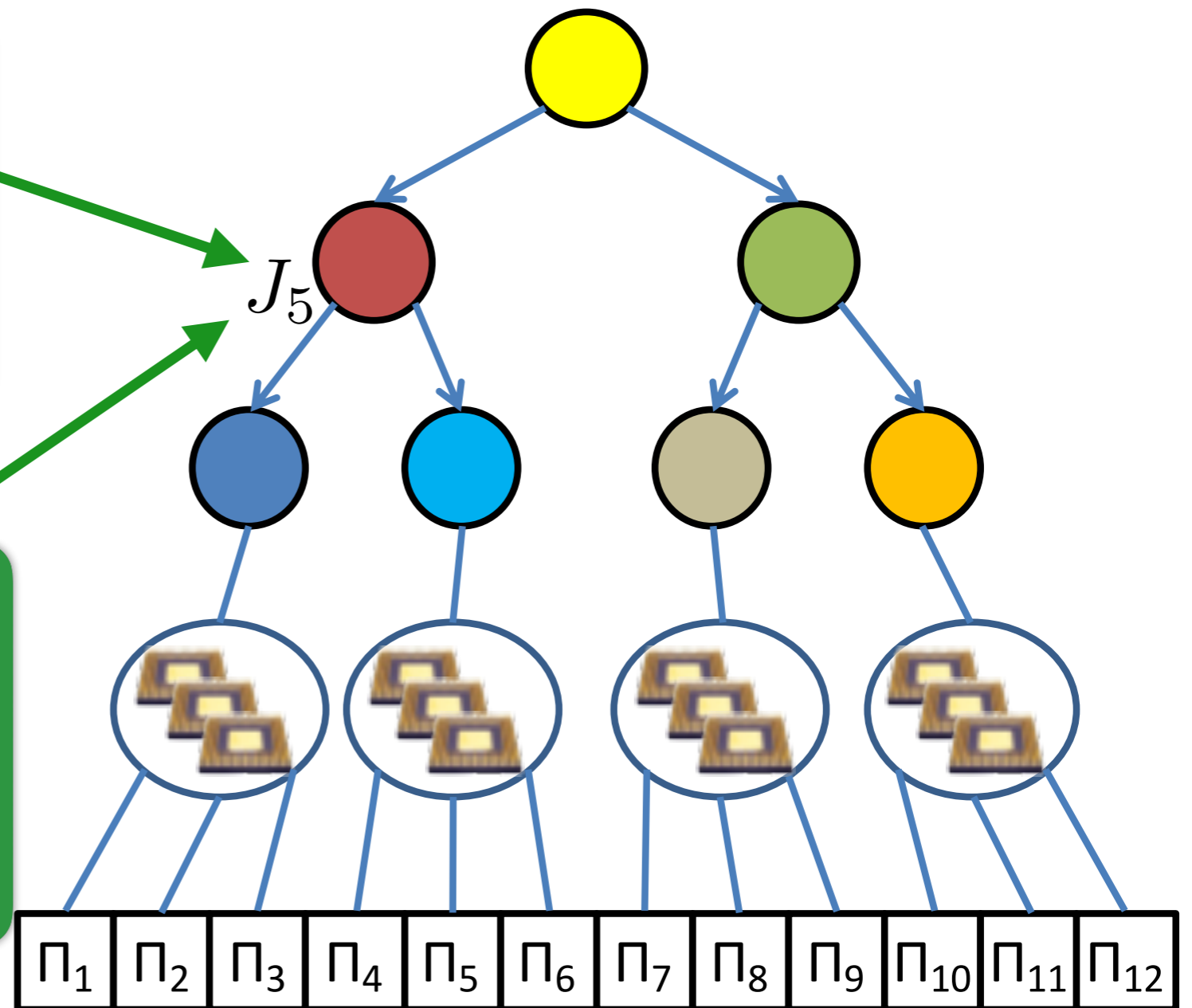
at most **one job** will effectively be **added to list of scheduled jobs**



Job Departure Step 4: Remove from Backlogged Heap

at most **one job** will effectively be **added to list of scheduled jobs**

remove this job from the heap of **backlogged jobs**



Complexity of Job

Departure: $O(\log n + m^2)$

n ...number of tasks

m ...number of cores

Complexity of Job

Departure: $O(\log n + m^2)$

1. Walk up the tree and remove departing job from lists: $O(\text{height of tree}) = O(m)$

n ...number of tasks

m ...number of cores

Complexity of Job Departure: $O(\log n + m^2)$

1. Walk up the tree and remove departing job from lists: $O(\textit{height of tree}) = O(m)$
2. Find highest-priority backlogged job in each affinity: $O(\textit{\#distinct affinities}) = O(m)$

n ...number of tasks

m ...number of cores

Complexity of Job Departure: $O(\log n + m^2)$

1. Walk up the tree and remove departing job from lists: $O(\textit{height of tree}) = O(m)$
2. Find highest-priority backlogged job in each affinity: $O(\textit{\#distinct affinities}) = O(m)$
3. Simulate arrivals: $O(\textit{\#distinct affinities} \times m) = O(m^2)$

n ...number of tasks

m ...number of cores

Complexity of Job Departure: $O(\log n + m^2)$

1. Walk up the tree and remove departing job from lists: $O(\textit{height of tree}) = O(m)$
2. Find highest-priority backlogged job in each affinity: $O(\textit{\#distinct affinities}) = O(m)$
3. Simulate arrivals: $O(\textit{\#distinct affinities} \times m) = O(m^2)$
4. Remove from backlogged: $O(\log n)$

n ...number of tasks

m ...number of cores

Speed-Up Bounds

Speed-up bound X for algorithm A

If a task set is schedulable **under *any policy*** on m **unit-speed processors**, then it is also schedulable under A with m ***processors of speed X*** .

- quantifiable relation to system **optimality**
- a way to structure the space of non-optimal algorithms
- the lower the speed-up bound, the better

First Speed-Up Results for Real-Time Scheduling with Affinity Restrictions

Considered special cases:

- job priorities determined with **EDF**

and either

- **bi-level** affinities or
- **clustered** affinities.

Bi-Level Affinities

- each task is assigned either
 - a **global** affinity (can use all cores) or
 - a **singleton** affinity (can use only one specific core)

Bi-Level Affinities

- each task is assigned either
 - a **global** affinity (can use all cores) or
 - a **singleton** affinity (can use only one specific core)

HPA-EDF + Bi-Level Affinities

required speed-up s : $s < 2.415$

Clustered Affinities

- each task is assigned either
 - a **global** affinity (can use all cores) or
 - a **clustered** affinity (can use only subset of cores)
 - all **clusters are mutually disjoint**

Clustered Affinities

- each task is assigned either
 - a **global** affinity (can use all cores) or
 - a **clustered** affinity (can use only subset of cores)
- all **clusters are mutually disjoint**

HPA-EDF + Clustered Affinities

required speed-up s : $s < 3.562$



Implementation in

LITMUSRT

Linux Testbed for Multiprocessor Scheduling in Real-Time Systems

www.litmus-rt.org

LITMUSRT

Linux Testbed for Multiprocessor Scheduling in Real-Time Systems

- real-time extension of the Linux kernel (*currently, Linux 4.1*)
- continuously maintained since 2006
- makes it ~~easy~~ **easier** to implement and evaluate (multiprocessor) real-time scheduling policies in Linux kernel on real hardware
- relevant highlights: built-in global **migration support** and **overhead tracing infrastructure**



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

[2006–2011]



Max
Planck
Institute
for
Software Systems

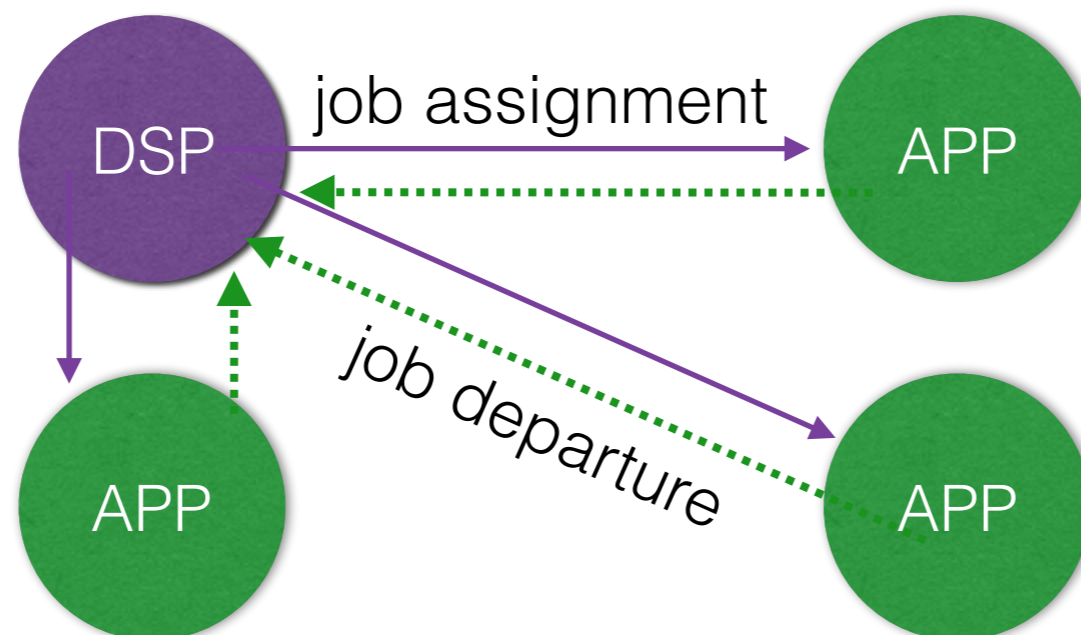
[2011–]

Evaluation Questions

- Can you actually run the proposed HPA scheduler **in a real OS kernel**?
- What **practical tweaks** are required?
- Isn't this algorithm prohibitively expensive in terms of **actual runtime overheads**?

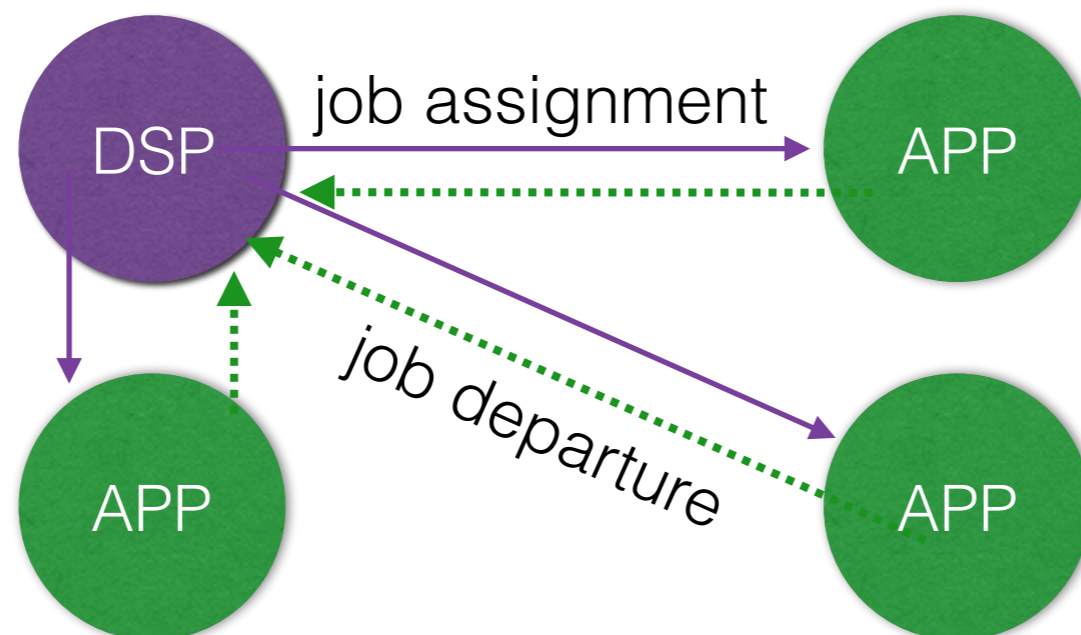
Baseline

- **HPA-FP** (HPA + fixed priority) implemented on top of Cerqueira et al.'s **message-passing-based global scheduler** [RTAS'14].
- **Basic idea**
 - one *designated scheduling processor* (**DSP**)
 - **DSP** makes *all* scheduling decisions (for all cores)
 - *application processors* send job state changes via messages
 - simple **dispatcher** enacts scheduling decisions on app procs.



Baseline

- **HPA-FP** (HPA + fixed priority) implemented on top of Cerqueira et al.'s **message-passing-based global scheduler** [RTAS'14].
- **Basic idea**
 - one *designated scheduling processor* (**DSP**)
 - **DSP** makes *all* scheduling decisions (for all cores)
 - *application processors* send job state changes via messages
 - simple **dispatcher** enacts scheduling decisions on app procs.



- + no locking of scheduler state
- + no cache-line bouncing
- + better scalability [max. overheads]

Practical Tweaks

Practical Tweaks

- **Affinity tree** is explicitly stored in kernel and **dynamically extended** as tasks are admitted

Practical Tweaks

- **Affinity tree** is explicitly stored in kernel and **dynamically extended** as tasks are admitted
- Message-passing-based design removes the need to synchronize tree traversals (big plus!).

Practical Tweaks

- **Affinity tree** is explicitly stored in kernel and **dynamically extended** as tasks are admitted
- Message-passing-based design removes the need to synchronize tree traversals (big plus!).
- **Strict Fibonacci heaps** are complicated & slow
 - use standard ***priority bitmap + linked lists***
 - ***effectively O(1)*** for fixed #priorities

Practical Tweaks

- **Affinity tree** is explicitly stored in kernel and **dynamically extended** as tasks are admitted
- Message-passing-based design removes the need to synchronize tree traversals (big plus!).
- **Strict Fibonacci heaps** are complicated & slow
 - use standard **priority bitmap + linked lists**
 - **effectively O(1)** for fixed #priorities
- **Locality-aware task mapping** to avoid needless migrations (Algorithm 6)
 - implemented with sets (=bit operations)
 - **effectively O(1)** for fixed, small #cores

Platform & Workloads

Platform

- Xeon E7 8857, two sockets, 12 cores each ($m = 24$)
- private **L1** and **L2** (32 KiB and 256 KiB, resp.)
- shared **L3** (30 MiB) per socket

Workload

- 75%/85% utilization
- execution costs: Emberson et al. (2010)
- log-uniform periods **1ms** to **1000ms**
- $2m$ to $10m$ tasks (48 to 240)
- three affinity levels: ***global, socket, partitioned***
- rate-monotonic priorities

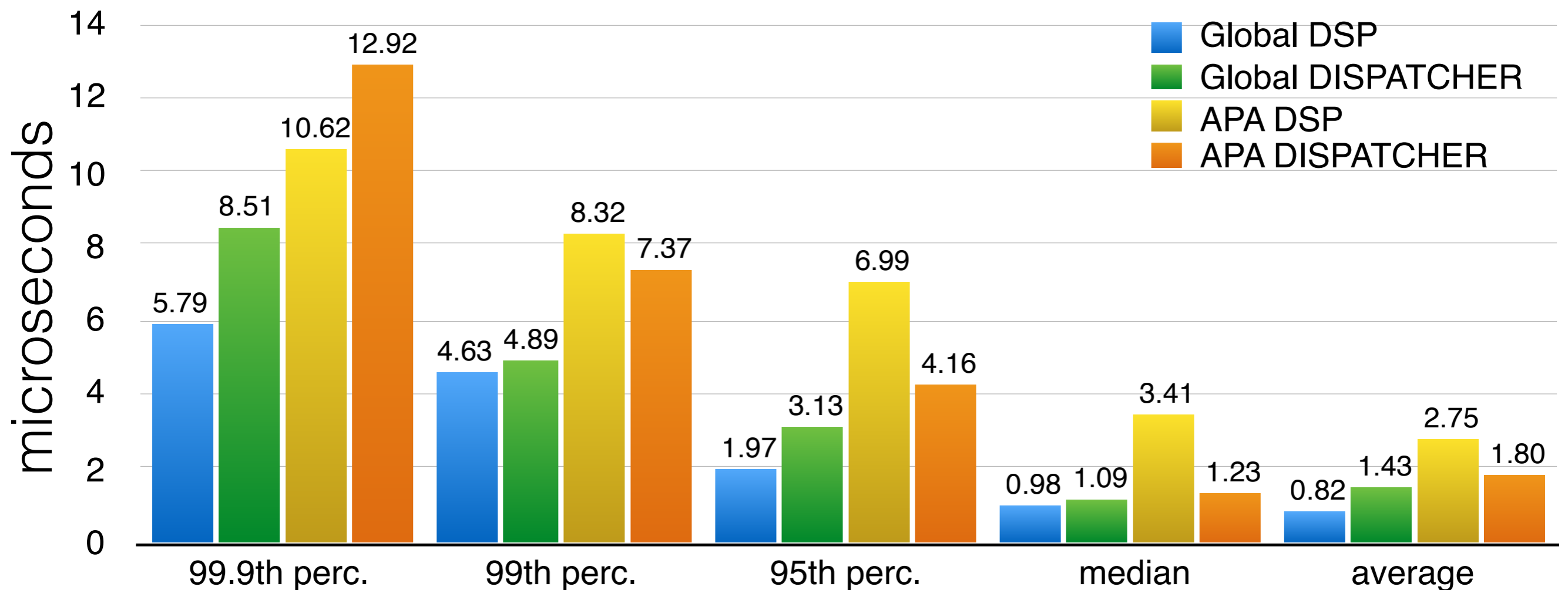
Experiments

- 150 task sets per scheduler
- 60 seconds per task set
- traced **scheduler overheads** with Feather-Trace
- 34 GiB trace data
- extracted 700,000,000 valid samples



Results Overview

- substantially increased costs ($\sim 1.5x$ to $\sim 3.5x$), but still in a feasible range (*a few microseconds*)



Evaluation Questions

- Can you actually run the proposed HPA scheduler **in a real OS kernel**?
- What **practical tweaks** are required?
- Isn't this algorithm prohibitively expensive in terms of **actual runtime overheads**?

Evaluation Questions

- Can you actually run the proposed HPA scheduler **in a real OS kernel**?

→ **Yes!**

- What **practical tweaks** are required?
- Isn't this algorithm prohibitively expensive in terms of **actual runtime overheads**?

Evaluation Questions

- Can you actually run the proposed HPA scheduler **in a real OS kernel?**

→ **Yes!**

- What **practical tweaks** are required?
→ *locality-aware assignment* and *simpler queues*
- Isn't this algorithm prohibitively expensive in terms of **actual runtime overheads?**

Evaluation Questions

- Can you actually run the proposed HPA scheduler **in a real OS kernel?**

→ **Yes!**

- What **practical tweaks** are required?

→ *locality-aware assignment* and *simpler queues*

- Isn't this algorithm prohibitively expensive in terms of **actual runtime overheads?**

→ *more costly, but not prohibitively so*

Concluding Remarks

Summary

Summary

Hierarchical processor affinities are an important special case.

Summary

Hierarchical processor affinities are an important special case.

The laminar affinity structure **allows for a much more efficient online scheduler.**

Summary

Hierarchical processor affinities are an important special case.

The laminar affinity structure **allows for a much more efficient online scheduler.**

first speed-up result for real-time scheduling with restricted processor affinities

Summary

Hierarchical processor affinities are an important special case.

The laminar affinity structure **allows for a much more efficient online scheduler.**

first speed-up result for real-time scheduling with restricted processor affinities

first implementation of a strong APA scheduler in a real OS kernel

Some Open Questions

- A more **efficient *weak* HPA** scheduler?
- **Speed-up** bounds for more **general cases**?
- More **accurate schedulability tests** for strong and weak HPA scheduling?
- Is there some **interesting class of affinities** between arbitrary and hierarchical?

APA > ?PA > HPA

LITMUS^{RT}

Linux Testbed for Multiprocessor Scheduling in Real-Time Systems

- New release **2016.1**
 - framework for proper **reservation-based scheduling**
- A new tutorial: **Getting Started with LITMUS^{RT}**
 - <http://www.litmus-rt.org/tutor16/>
- Detailed **artifact evaluation** instructions
 - how to run our HPA scheduler
 - how to collect and process data
 - <https://www.mpi-sws.org/~bbb/papers/ae/ecrts16/laminar-apa.html>

