(Overview of Cooper vs. Hera analysis)

Below, we discuss (a) the performance penalty among all possible pair of co-located models, which is used by Cooper to (b) evaluate model "preference" as defined in the Cooper paper. The evaluated model preference is used to (c) determine the best pair of models to co-locate, which we utilized to (d) compare the aggregate QPS achieved when employing Cooper and Hera's model selection algorithm. Because Cooper strictly focuses on efficient model selection policies without resource partitioning, we employ Hera's resource management algorithm for "both" Cooper and Hera.

(Co-located model)

(Target model)	DLRM (A)	DLRM (B)	DLRM (C)	DLRM (D)	NCF	DIEN	DIN	WnD
DLRM (A)	-	30	20	20	20	20	20	10
DLRM (B)	10	-	40	40	60	0	0	10
DLRM (C)	10	40	-	20	60	20	10	20
DLRM (D)	40	60	10	-	40	10	0	10
NCF	30	60	30	10	-	30	30	40
DIEN	0	0	0	0	40	-	0	0
DIN	0	20	0	0	0	10	-	0
WnD	30	40	40	20	20	30	30	-

(a) Performance penalty (%) when a given model (gray) gets co-located with another model (black). For instance, DLRM(A) experiences 30% performance penalty when co-located with DLRM(B). All experiments assume Hera's resource management unit (RMU) is employed for intra-node resource management.

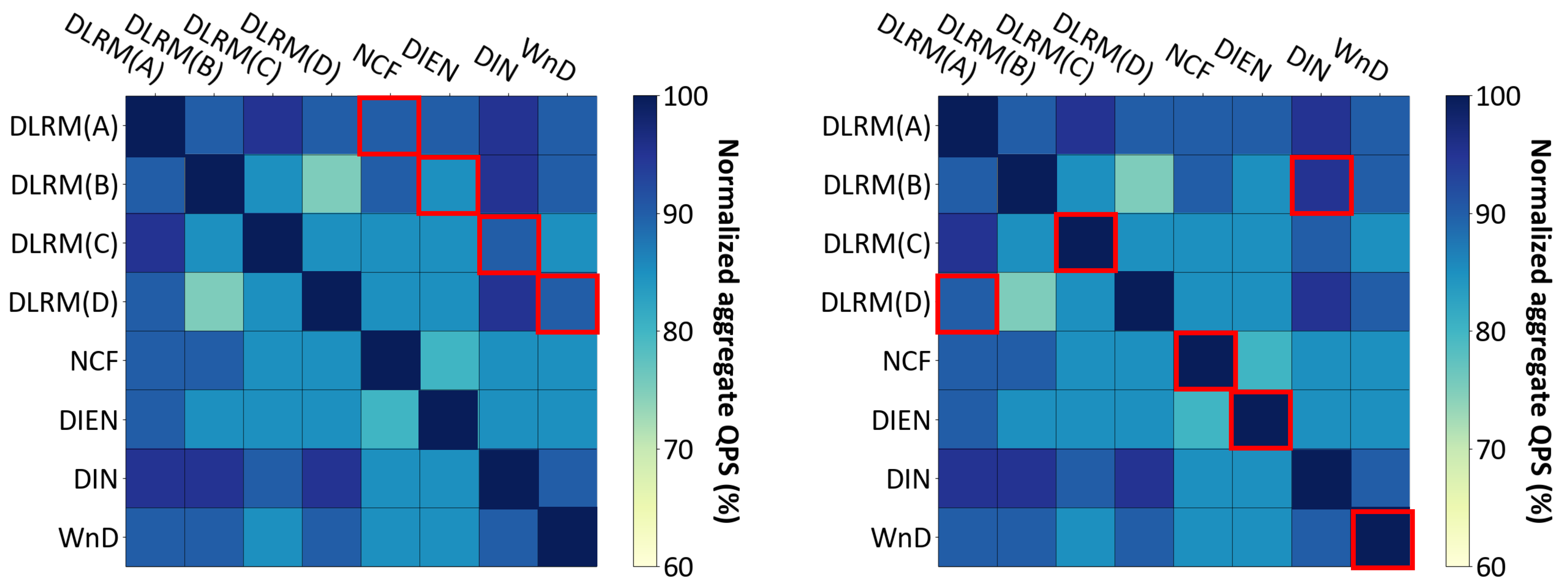
High preference → Low preference

(Target model)							
DLRM (A)	WnD	DLRM (C)	NCF	DIEN	DIN	DLRM (D)	DLRM (B)
DLRM (B)	DIEN	DIN	DLRM(A)	WnD	DLRM (C)	DLRM (D)	NCF
DLRM (C)	DLRM(A)	DIN	DIEN	WnD	DLRM (D)	DLRM (B)	NCF
DLRM (D)	DIN	DLRM (C)	DIEN	WnD	DLRM(A)	NCF	DLRM (B)
NCF	DLRM (D)	DLRM (A)	DLRM (C)	DIEN	DIN	WnD	DLRM (B)
DIEN	DLRM(A)	DLRM (B)	DLRM (C)	DIN	WnD	DLRM (D)	NCF
DIN	DLRM(A)	DLRM (C)	NCF	WnD	DLRM (D)	DIEN	DLRM (B)
WnD	NCF	DLRM (D)	DLRM (A)	DIEN	DIN	DLRM (B)	DLRM (C)

(b) Model preference, sorted from high (left) to low (right) for each model. As defined in Cooper, a given model prefers other models causing less penalty when co-located together. For instance, DLRM(A) prefers WnD (experiences 10% penalty as shown in Figure(a) above) over other models (e.g., DLRM(C) which experiences 20% penalty) as it causes the least performance degradation under co-location.

DLRM (A)	NCF	DLRM (B)	DIEN	DLRM (C)	DIN	DLRM (D)	WnD

(c) The co-located model pairs determined by Cooper's stable roommate algorithm, i.e., one of the fairness-oriented partitioning algorithm proposed in Cooper. The pair of color-coded RecSys models shown in Figure(b) designate the finally chosen co-location pairs which are concentrated on the left-side of the preference scores, demonstrating Cooper's algorithmic tendency to pair models with high preferences among each other.



Cooper
Average normalized aggregate QPS: 88.75

Hera
Average normalized aggregate QPS: 97.5

(d) The RED boxes in each figure illustrates the selected model pairs for co-location when using Cooper (left) and Hera (right), e.g., Cooper decides to co-locate {DLRM(A)+NCF}, {DLRM(B)+DIEN}, {DLRM(C)+DIN}, and {DLRM(D)+WnD} as shown in Figure(c). Notice how the colors within each RED boxes in Cooper are generally lighter than those under Hera, designating lower aggregate QPS when employing Cooper.