??
$_t axonomy.pdf Taxonomy of layer partitioning schemes across PEs. (a) Batch partitioning divides different images across P$
Batch Partitioning (Data-parallelism):
?
Fmap (Image) Partitioning:
112×112
??
?
Output Partitioning:
K
p
K/p
Input Partitioning:
C
Importantly, input partitioning requires an All-Reduce operation to aggregate partial sums across PEs
??

Scheme	Weight	IFmap	OFmap	Reduction
			All-Reduce	Yes

?
hybrid partitioning scheme
fmap partitioning
output partitioning

$$E_{access} = A_{DRAM} \times e \times (1 + \beta \cdot r) \quad (1)$$

e
β
r
A_{DRAM}
r
A_{DRAM}
A_{DRAM}
r
INPP All-Reduce Cost:

$$E_{INPP} = D_{output} \times \frac{2(p-1)}{p} \times e_{comm} \quad (2)$$

D