Expert Deduplication Benefits Analysis Memory Savings vs Batch Size Bandwidth Savings vs Locality Factor 70 + 2 experts/item 4 experts/item 8 experts/item 60 16 experts/item 50 60 Bandwidth Savings (%) 8 0 0 Memory Savings (%) Batch 1 Batch 4 Batch 16 Batch 64 20 20 10 0 2⁴ 2⁵ 26 2^{1} 2² 0.2 0.3 0.6 0.4 0.1 0.5 0.7 Locality Factor Batch Size **Expert Reuse Factor** Memory vs Bandwidth Savings Batch 1 0.000 0.000 0.000 0.000 Batch 2 - 0.8 Batch 4 80 Batch 8 - 0.7 Batch 16 0.031 0.123 0.013 0.062 Batch 32 Batch 64 - 0.6 Bandwidth Savings (%) 0.075 0.147 0.205 0.259 Batch Size 8 ' - 0.5 0.198 0.276 0.408 0.094 - 0.4 16 0.133 0.263 0.408 0.587 - 0.3 20 - 0.2 0.223 0.402 0.583 0.757 - 0.1 64 0.377 0.584 0.759 0.875 0 - 0.0 2 8 16 20 40 60 80 Experts per Item Memory Savings (%)