Latency vs GPU Heterogeneity

Seq Len = 4,096 · Seq Len = 8,192 · Seq Len = 16,384 · Seq Len = 32,768 · Seq Len = 65,536

Even Split · Uneven Split · LUT Split · Formula Split · Homogeneous