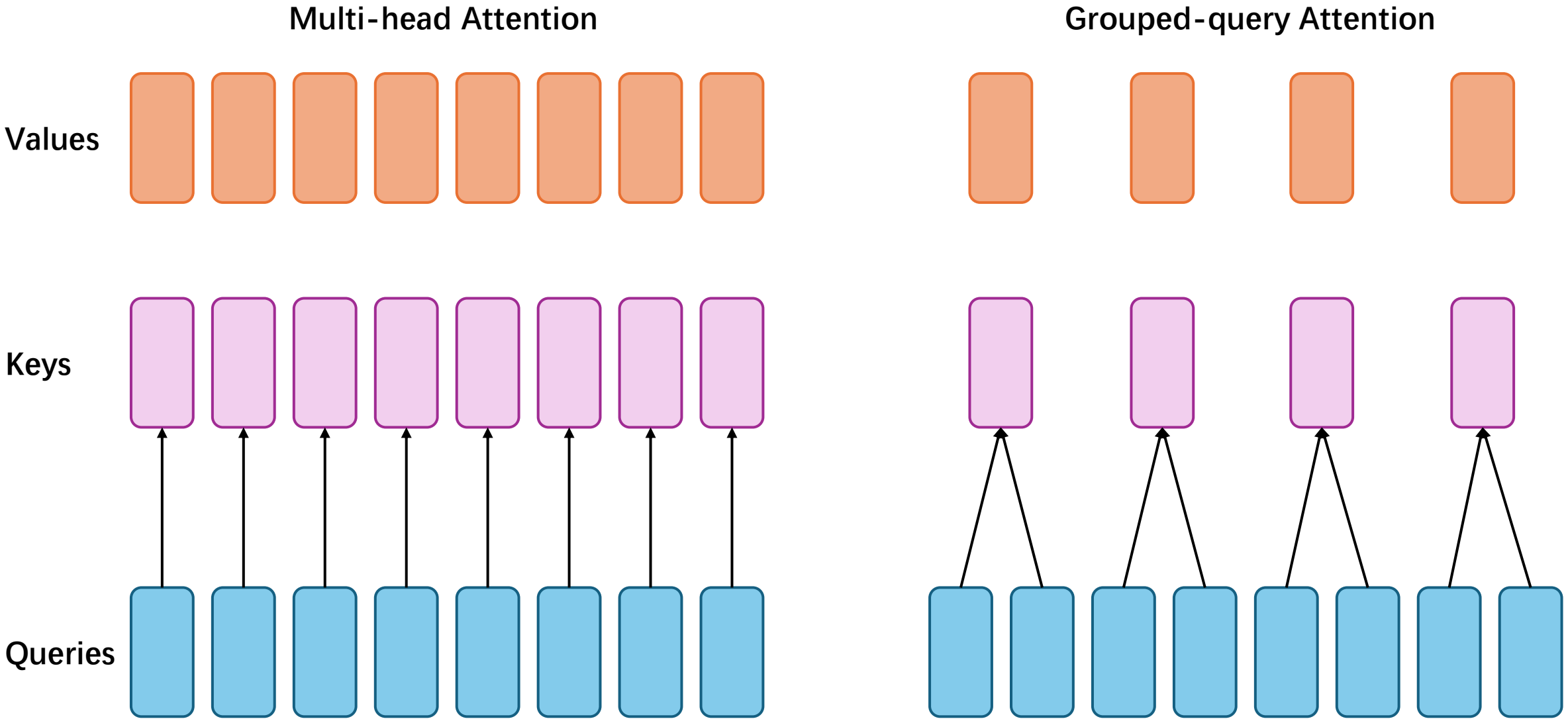
Multi-head Attention
Grouped-query Attention
Values
Keys
Queries