A Supplementary Material

A.1 Backward pass for ZeCO with All-Scan comunication

In the backward propagation of the ZeCO algorithm, most of the process is similar to the forward propagation. It is important to note the difference in notation here: $\tilde{\boldsymbol{\gamma}}_{[n]}$ denotes the decay factor for the reverse cumulative product. Furthermore, in the official implementation of gated linear attention, $\mathbf{S}_{[n]}$ needs to be recomputed during the backward pass. However, since the global initial state has already been obtained during the forward pass, there is no need for all-scan communication when recomputing $\mathbf{S}_{[n]}$.

Algorithm 1 Backward pass for ZeCO with All-Scan comunication

Input: $\mathbf{Q}, \mathbf{K}, \mathbf{G} \in \mathbb{R}^{L \times d_k}, \mathbf{V}, \mathbf{dO} \in \mathbb{R}^{L \times d_v}$, chunk size C, num_device P, device_rank $p \in \{0, 1, \ldots, P-1\}$

Initialize $\mathbf{dS} = \mathbf{0} \in \mathbb{R}^{d_k \times d_v}$ on SRAM

for $n \leftarrow N$ **to** 0 **do**
  Load $\mathbf{G}_{[n]} \in \mathbb{R}^{C \times d_k}$ from HBM to SRAM
  Load $\mathbf{Q}_{[n]} \in \mathbb{R}^{C \times d_k}$ from HBM to SRAM
  Load $\mathbf{dO}_{[n]} \in \mathbb{R}^{C \times d_v}$ from HBM to SRAM
  On chip, compute $\boldsymbol{\gamma}_{[n]}, \boldsymbol{\Gamma}_{[n]}$ and $\tilde{\mathbf{Q}}_{[n]} = \mathbf{Q}_{[n]} \odot \mathbf{G}_{[n]}, \tilde{\gamma} = \tilde{\gamma} \odot \boldsymbol{\gamma}_{[n]}$
  Store $\tilde{\gamma}$ in HBM as $\tilde{\gamma}_{[n]}$
  On chip, compute $\mathbf{dS} = (\boldsymbol{\gamma}_{[n]}^\top \mathbf{1}) \odot \mathbf{dS} + \tilde{\mathbf{Q}}_{[n]}^\top \mathbf{dO}_{[n]}$
  Store $\mathbf{dS}$ in HBM as $\mathtt{dS}_{[n]}$
end for
In parallel do:
parallel stream 1:
$\mathbf{dS}_{(p-1)L}, \mathbf{dS}_{pL} \leftarrow$ All-Scan$(\mathbf{dS}_{[0]}, \tilde{\gamma}_{[0]})$
parallel stream 2:
Load $\mathbf{S}_{(p-1)L}$ from HBM to SRAM
On chip, recompute $\mathbf{S}_{[n]}$ with $\mathbf{S}_{[0]} = \mathbf{S}_{(p-1)L}, n = \{0, 1, 2, \ldots, N-1\}$
Store $\{\mathbf{S}_{[n]}, n \in \{0, 1, 2, \ldots, N-1\}\}$
for $n \leftarrow 1$ **to** N **in parallel do**
  Load $\mathbf{Q}_{[n]}, \mathbf{K}_{[n]}, \mathbf{G}_{[n]}, \mathbf{V}_{[n]}, \mathbf{dO}_{[n]}$ from HBM to SRAM
  Load , $\in \mathbb{R}^{d_k \times d_v}$, from HBM to SRAM
  On chip, construct causal mask $\mathbf{M} \in \mathbb{R}^{B \times B}$
  On chip, compute $\boldsymbol{\Lambda}_{[n]}, \boldsymbol{\Gamma}_{[n]} \in \mathbb{R}^{C \times d_k}$
  On chip, compute $\tilde{\mathbf{Q}}_{[n]} = \mathbf{Q}_{[n]} \odot \boldsymbol{\Lambda}_{[n]}, \tilde{\mathbf{K}}_{[n]} = \mathbf{K}_{[n]} \odot \boldsymbol{\Gamma}_{[n]}$
  On chip, compute $\mathbf{P}_{[n]} = (\tilde{\mathbf{Q}}_{[n]} \tilde{\mathbf{K}}_{[n]}^\top) \odot \mathbf{M} \in \mathbb{R}^{C \times C}$
  On chip, compute $\mathbf{dP}_{[n]} = (\mathbf{dO}_{[n]} \mathbf{V}_{[n]}^\top) \odot \mathbf{M}$
  On chip, compute $\mathbf{d}\tilde{\mathbf{K}}_{[n]} = \tilde{\mathbf{Q}}_{[n]}^\top \mathbf{dP}$
  On chip, compute $\mathbf{dK}_{[n]} = \mathbf{d}\tilde{\mathbf{K}}_{[n]} / \boldsymbol{\Lambda}_{[n]}$
  On chip, compute $\mathbf{d}\tilde{\mathbf{Q}}_{[n]} = \mathbf{dP} \tilde{\mathbf{K}}_{[n]}$
  On chip, compute $\mathbf{dQ}_{[n]} = \mathbf{d}\tilde{\mathbf{Q}}_{[n]} \odot \boldsymbol{\Lambda}_{[n]}$
  Store $\mathbf{P}_{[n]}, \mathbf{dQ}_{[n]}, \mathbf{dK}_{[n]}$ in HBM.
end for
stream barrier
for $n \leftarrow 1$ **to** N **in parallel do**
  Load $\mathbf{P}_{[n]}, \mathbf{dQ}_{[n]}, \mathbf{dK}_{[n]}, \mathbf{dO}_{[n]}, \mathbf{Q}_{[n]}, \mathbf{K}_{[n]}, \mathbf{G}_{[n]}, \tilde{\gamma}_{[n-1]}, \mathbf{dS}_{pL}, \mathbf{S}_{[n-1]}$, from HBM to SRAM
  On chip, compute $\boldsymbol{\Lambda}_{[n]}, \boldsymbol{\Gamma}_{[n]} \in \mathbb{R}^{C \times d_k}$
  On chip, compute $\tilde{\mathbf{K}}_{[n]} = \mathbf{K}_{[n]} \odot \boldsymbol{\Gamma}_{[n]}$
  On chip, compute $\mathbf{d}\tilde{\mathbf{K}}_{[n]} = \mathbf{V}_{[n]} (\mathbf{dS}_{[n-1]}^\top + (\tilde{\gamma}_{[n-1]}^\top \mathbf{1}) \odot \mathbf{dS}_{pL}^\top)$
  On chip, compute $\mathbf{dK}_{[n]} = \mathbf{dK}_{[n]} + \mathbf{d}\tilde{\mathbf{K}}_{[n]} \odot \boldsymbol{\Gamma}_{[n]}$
  On chip, compute $\mathbf{d}\tilde{\mathbf{Q}}_{[n]} = \mathbf{dO}_{[n]} \mathbf{S}_{[n-1]}^\top$
  On chip, compute $\mathbf{dQ}_{[n]} = \mathbf{dQ}_{[n]} + \mathbf{d}\tilde{\mathbf{Q}}_{[n]} \odot \boldsymbol{\Lambda}_{[n]}$
  On chip, compute $\mathbf{dV}_{[n]} = \mathbf{P}_{[n]}^\top \mathbf{dO}_{[n]} + \tilde{\mathbf{K}}_{[n]} (\mathbf{dS}_{[n-1]}^\top + (\tilde{\gamma}_{[n-1]}^\top \mathbf{1}) \odot \mathbf{dS}_{pL}^\top)$
  Store $\mathbf{dK}_{[n]}, \mathbf{dV}_{[n]}$ in HBM
end for
Let $\mathbf{dQ} = \{\mathbf{dQ}_{[1]}, \ldots, \mathbf{dQ}_{[N]}\}, \mathbf{dK} = \{\mathbf{dK}_{[1]}, \ldots, \mathbf{dK}_{[N]}\}, \mathbf{dV} = \{\mathbf{dV}_{[1]}, \ldots, \mathbf{dV}_{[N]}\}$
Compute $\mathbf{dA} = \mathbf{Q} \odot \mathbf{dQ} - \mathbf{K} \odot \mathbf{dK}, \mathbf{dG} = \mathtt{revcum}(\mathbf{dA})$
return $\mathbf{dQ}, \mathbf{dK}, \mathbf{dV}, \mathbf{dG}$