Weight Reuse for LSTM networks

LSTM Equations

$$\tilde{c}^{<t>} = tanh(W_{x_c} x^{<t>} + W_{h_c} h^{<t-1>} + b_c)$$

$$u^{<t>} = \sigma(W_{x_u} x^{<t>} + W_{h_u} h^{<t-1>} + b_u)$$

$$f^{<t>} = \sigma(W_{x_f} x^{<t>} + W_{h_f} h^{<t-1>} + b_f)$$

$$o^{<t>} = \sigma(W_{x_o} x^{<t>} + W_{h_o} h^{<t-1>} + b_o)$$

Gates

LSTM Equations

$$\tilde{c}^{<t>} = tanh(W_{x_c} x^{<t>} + W_{h_c} h^{<t-1>} + b_c)$$

$$u^{<t>} = \sigma(W_{x_u} x^{<t>} + W_{h_u} h^{<t-1>} + b_u)$$

$$f^{<t>} = \sigma(W_{x_f} x^{<t>} + W_{h_f} h^{<t-1>} + b_f)$$

$$o^{<t>} = \sigma(W_{x_o} x^{<t>} + W_{h_o} h^{<t-1>} + b_o)$$

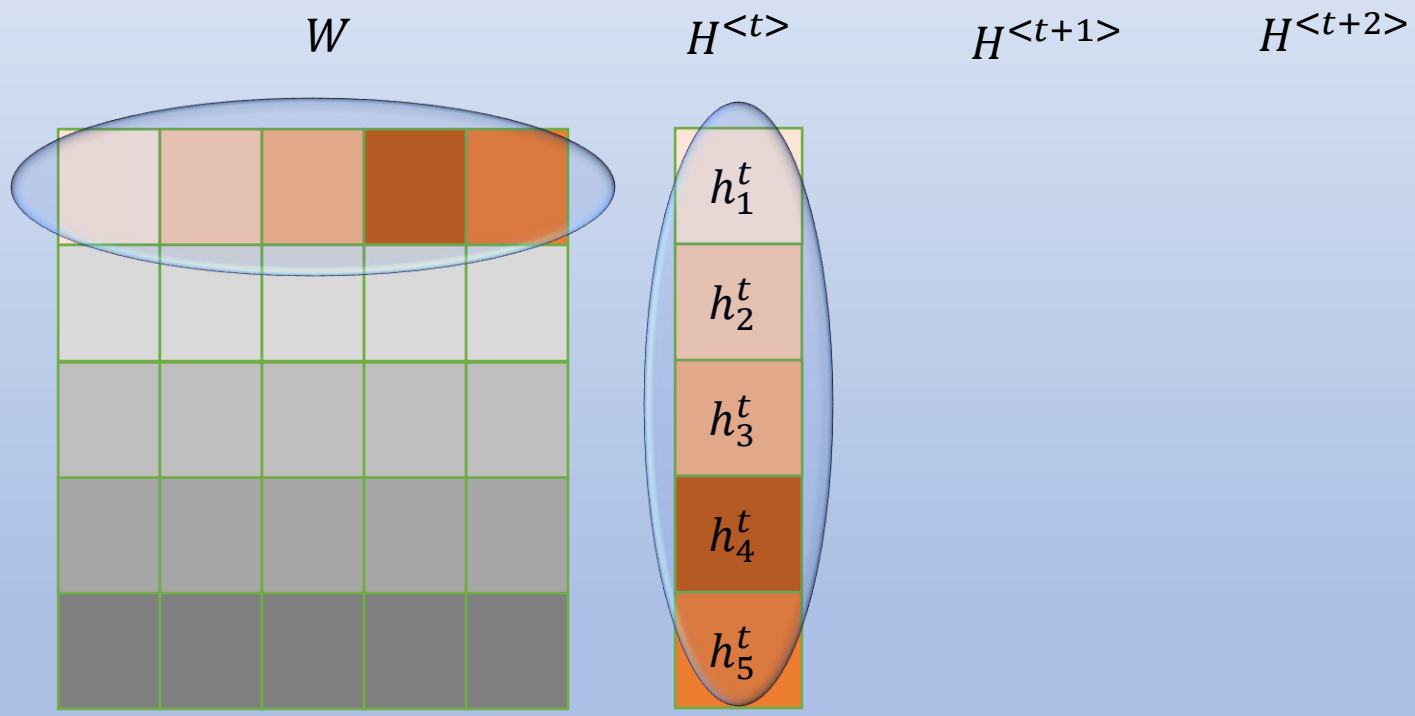
Gates (Matrix-Vector Mult)

$$c^{<t>} = (u^{<t>} \odot \tilde{c}^{<t>} + f^{<t>} \odot \tilde{c}^{<t>})$$

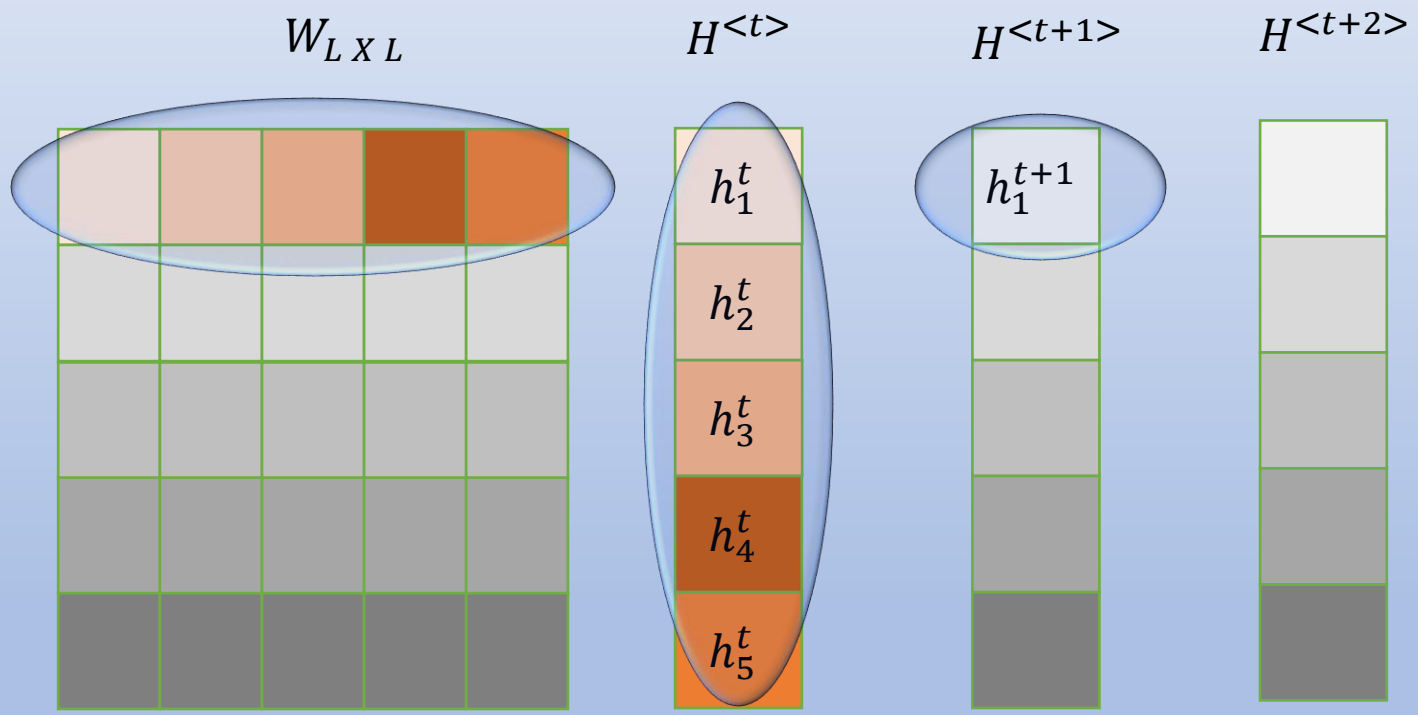
$$h^{<t>} = tanh(o^{<t>} \odot c^{<t>})$$

Update (Vector-Vector Mult

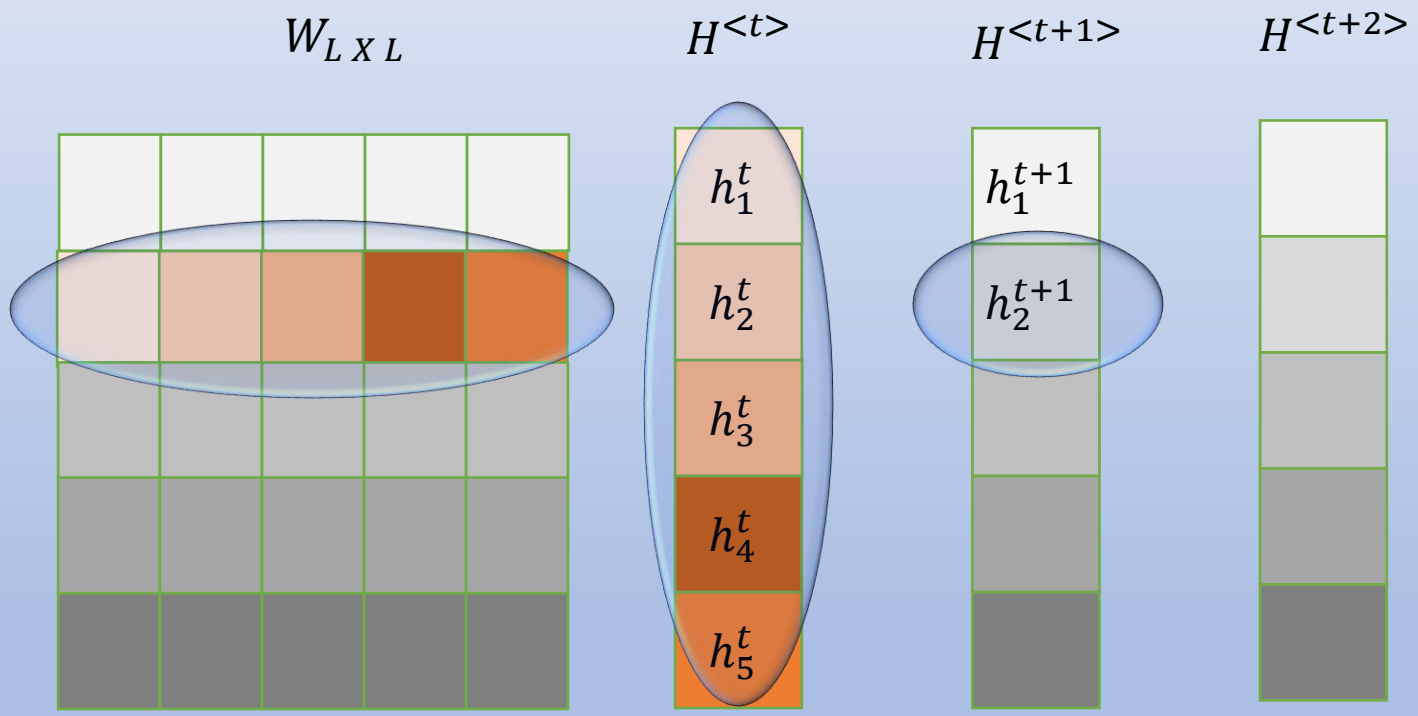
Matrix Vector Multiplication Dependency



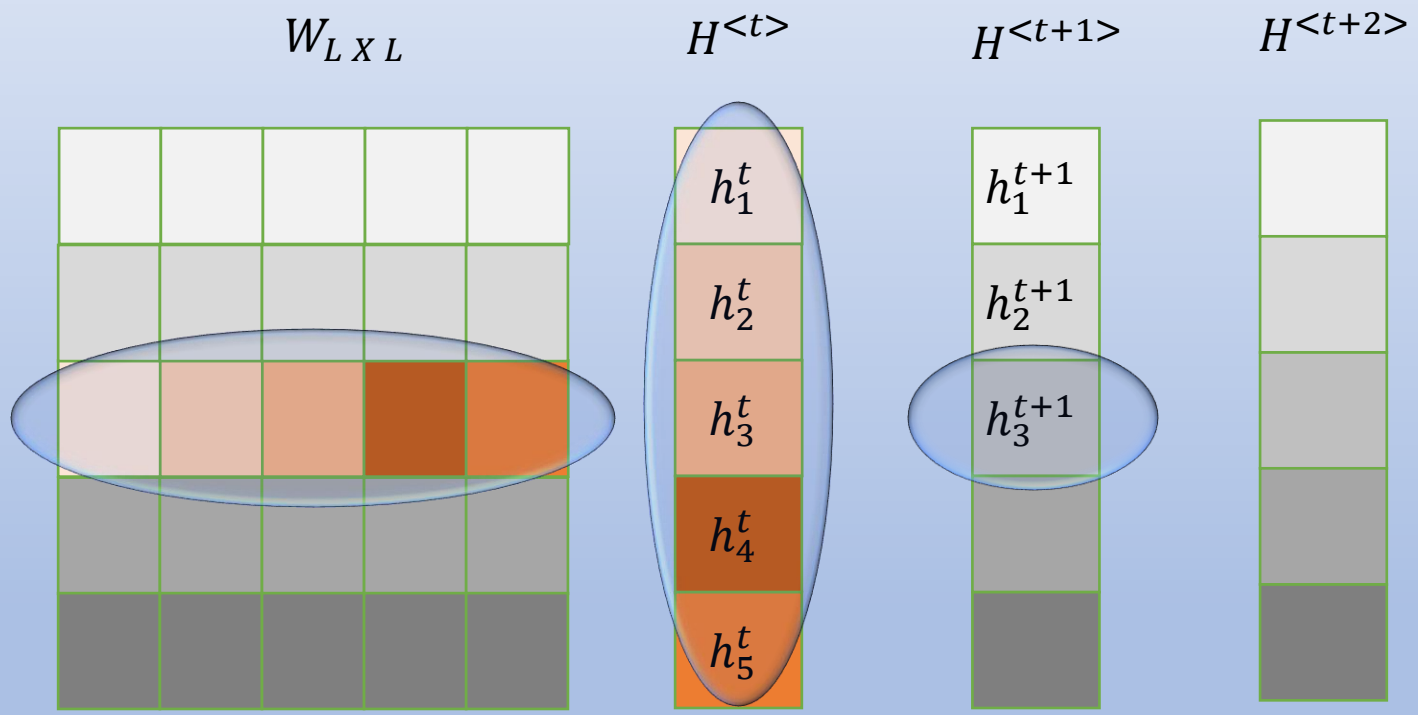
Matrix Vector Multiplication Dependency



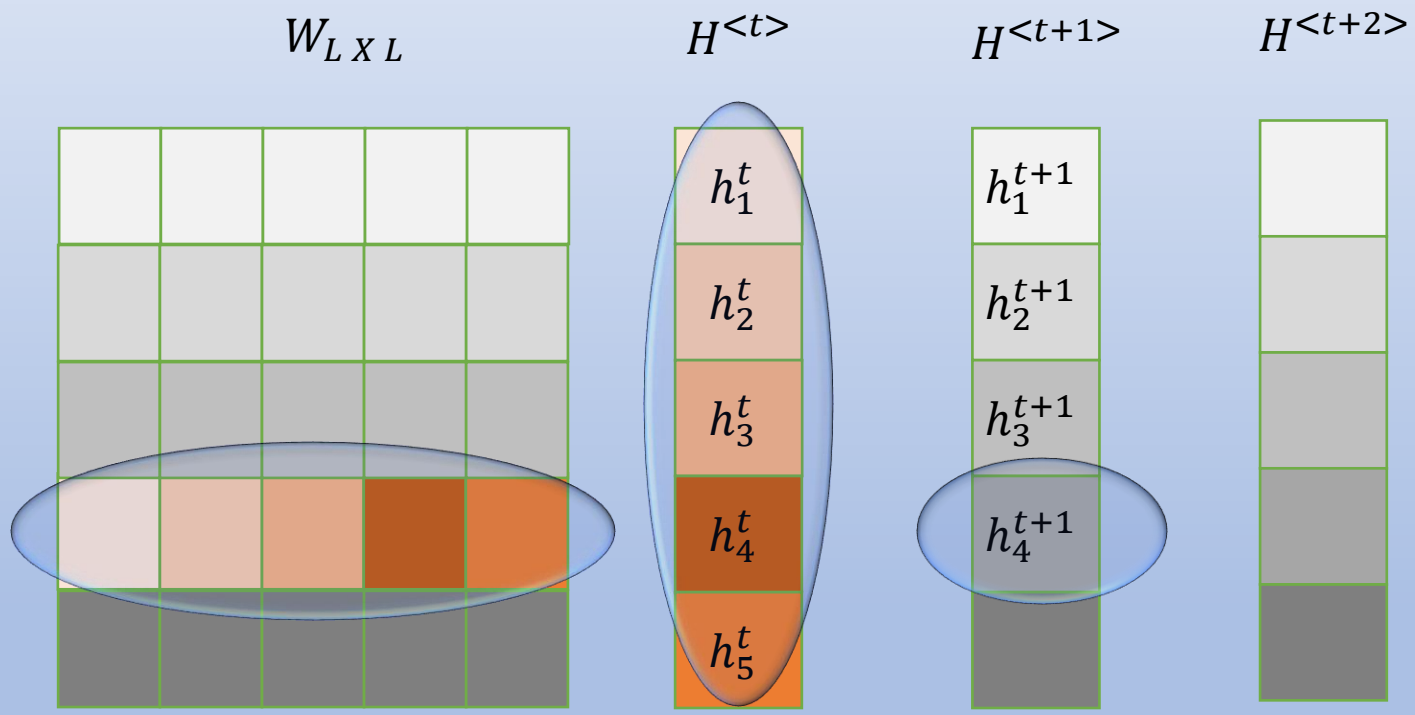
Matrix Vector Multiplication Dependency



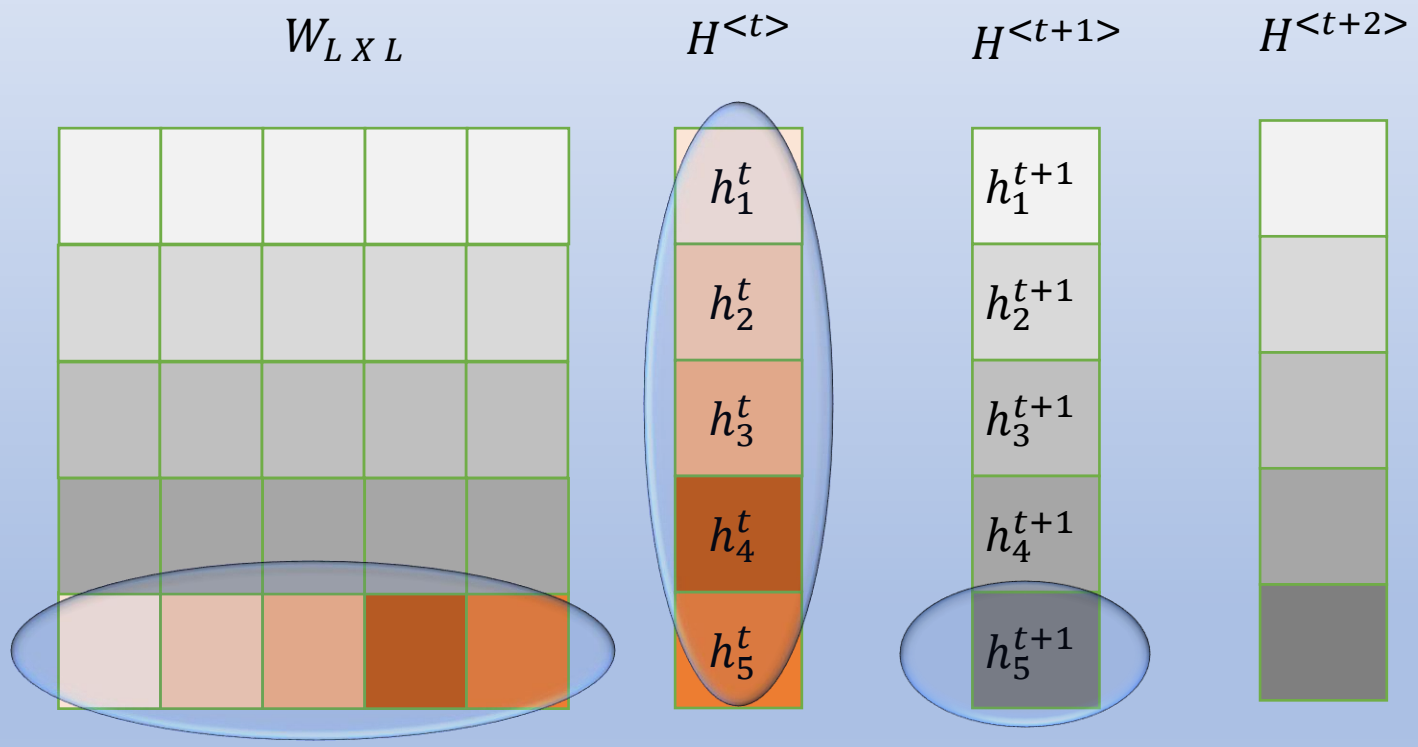
Matrix Vector Multiplication Dependency



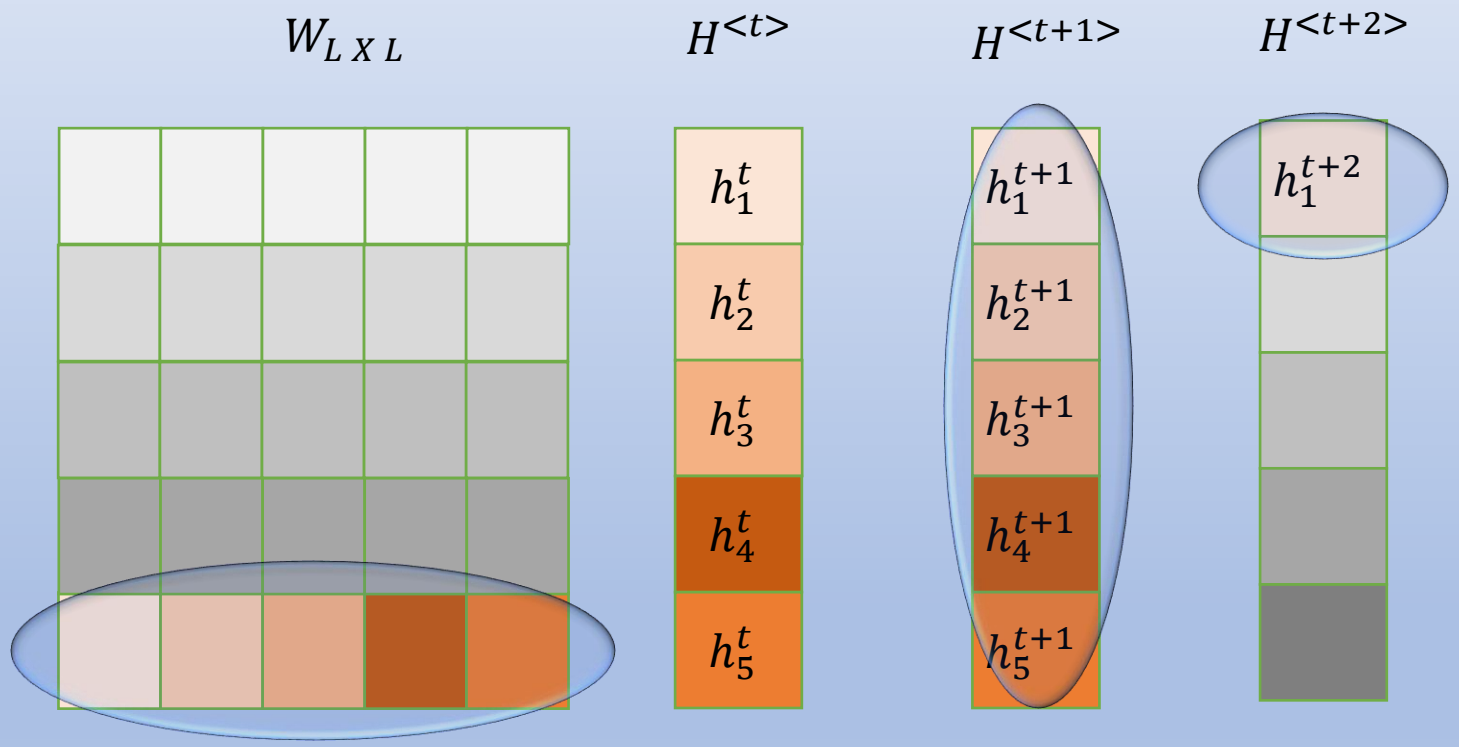
Matrix Vector Multiplication Dependency



Matrix Vector Multiplication Dependency



Matrix Vector Multiplication Dependency



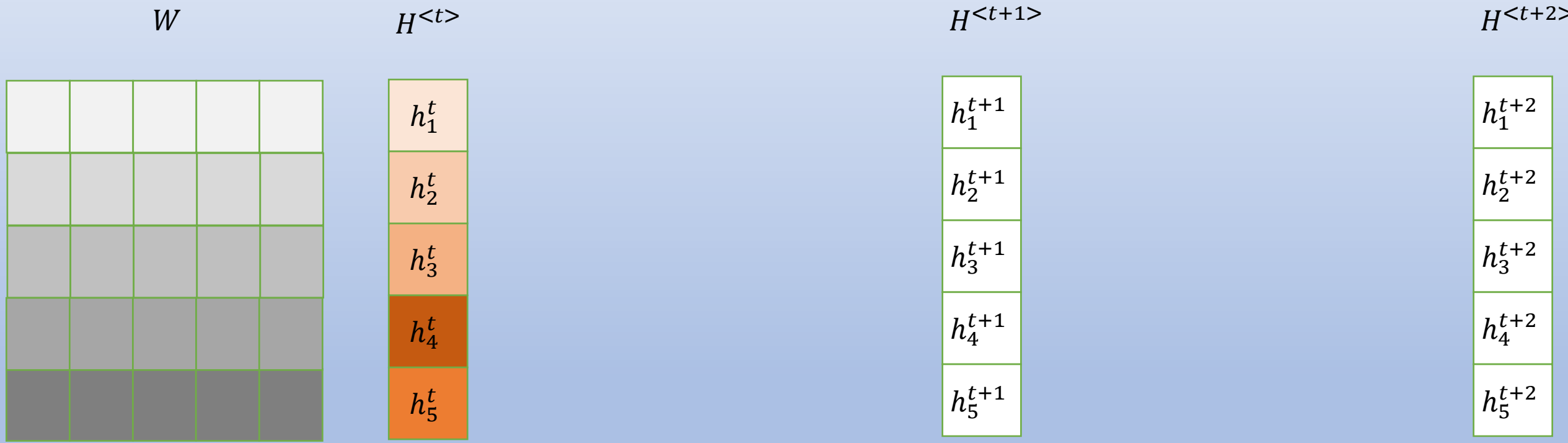
Off-Chip Memory Access

- W_h matrices are re-used at each time step.
- W_h is a square matrix of size 128x128x2 (=32KB), 256x256x2 (=128KB),or 512x512x2(=512KB).
- Dependencies $H^{<t+1>} \rightarrow H^{<t+1>} \rightarrow H^{<t+2>}$ limits the data reuse
- Due to limited on-chip memory , all 4 W_h matrices need to be accessed from off-chip memory at each <t>.
  - Large volume of off-chip memory access.
  - High Energy consumption and
  - Latency.

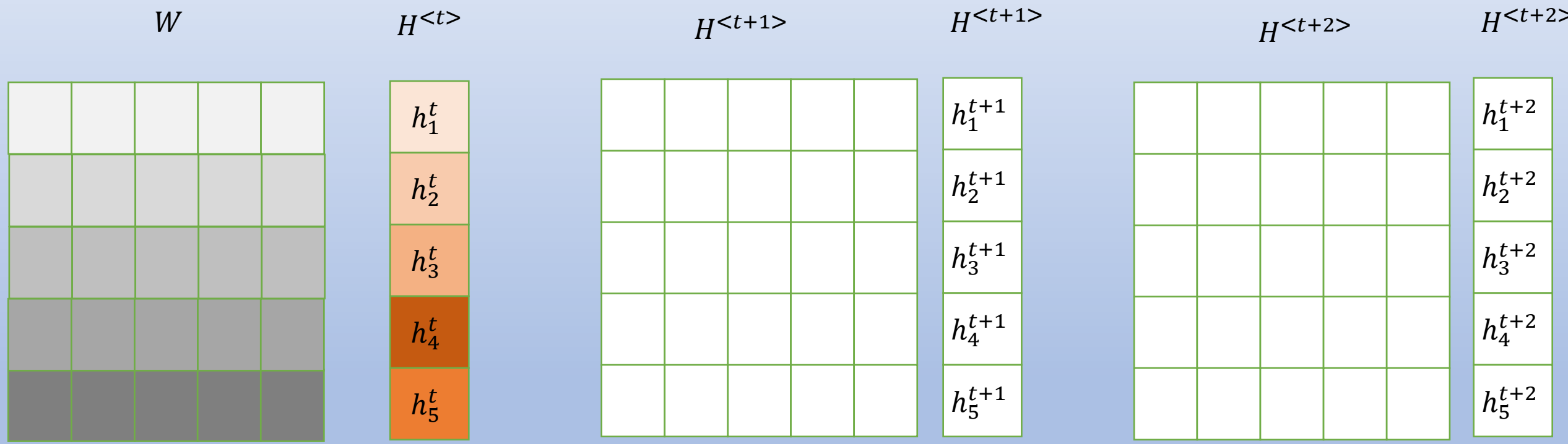
Previous Work

- J. Park, W. Yi, D. Ahn, J. Kung and J. -J. Kim, "Balancing Computation Loads and Optimizing Input Vector Loading in LSTM Accelerators," 2019, in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.
- Z. Que et al., "Efficient Weight Reuse for Large LSTMs," 2019, ASAP.
- Naebeom Park, Yulhwa Kim, Daehyun Ahn, Taesu Kim, and Jae-Joon Kim. 2020. “Time-step interleaved weight reuse for LSTM neural network computing”. ISLPED ‘2020.

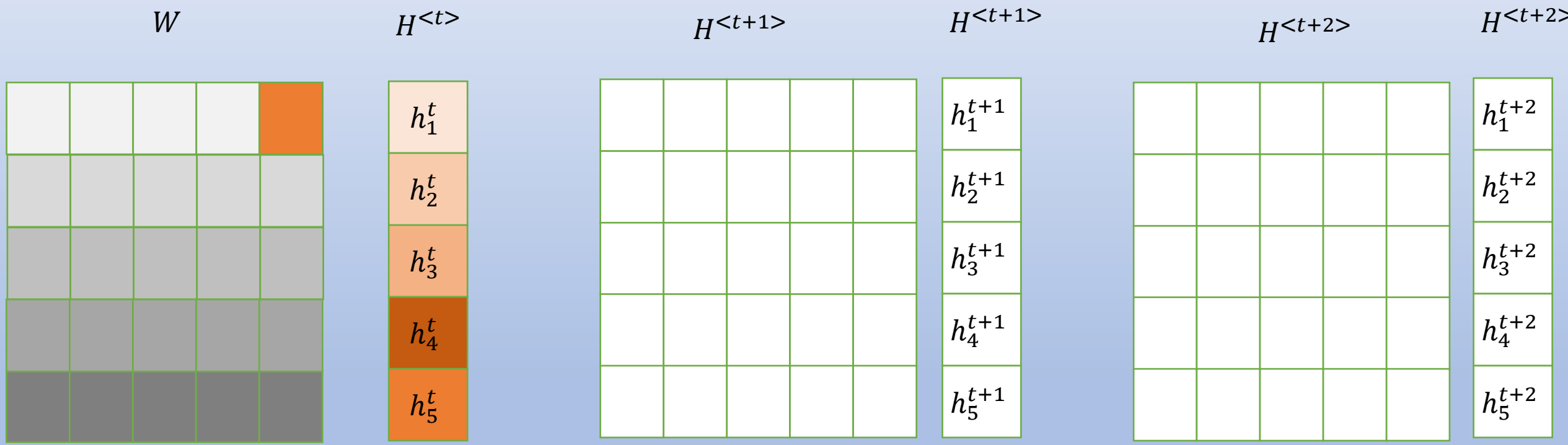
Proposed Approach: Weight Matrix Data Reuse Over Time



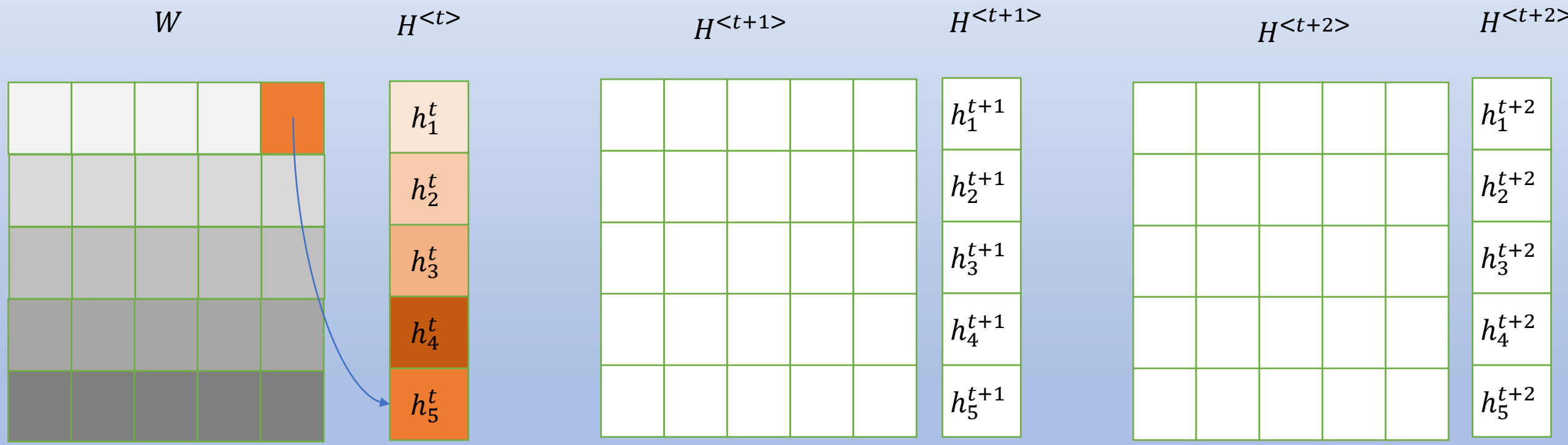
Proposed Approach: Weight Matrix Data Reuse Over Time



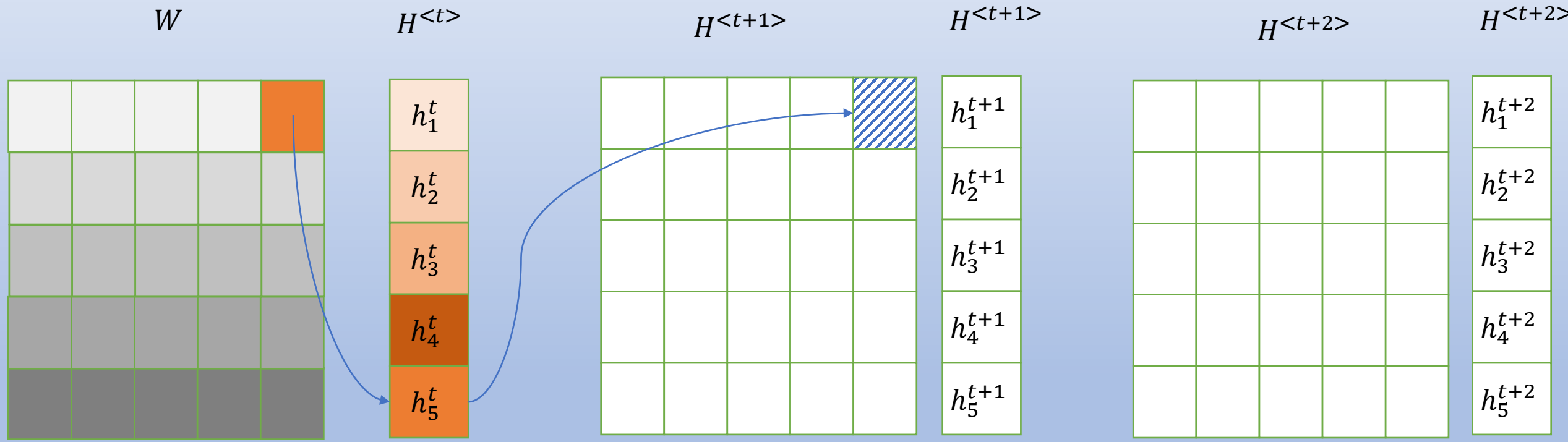
Proposed Approach: Weight Matrix Data Reuse Over Time



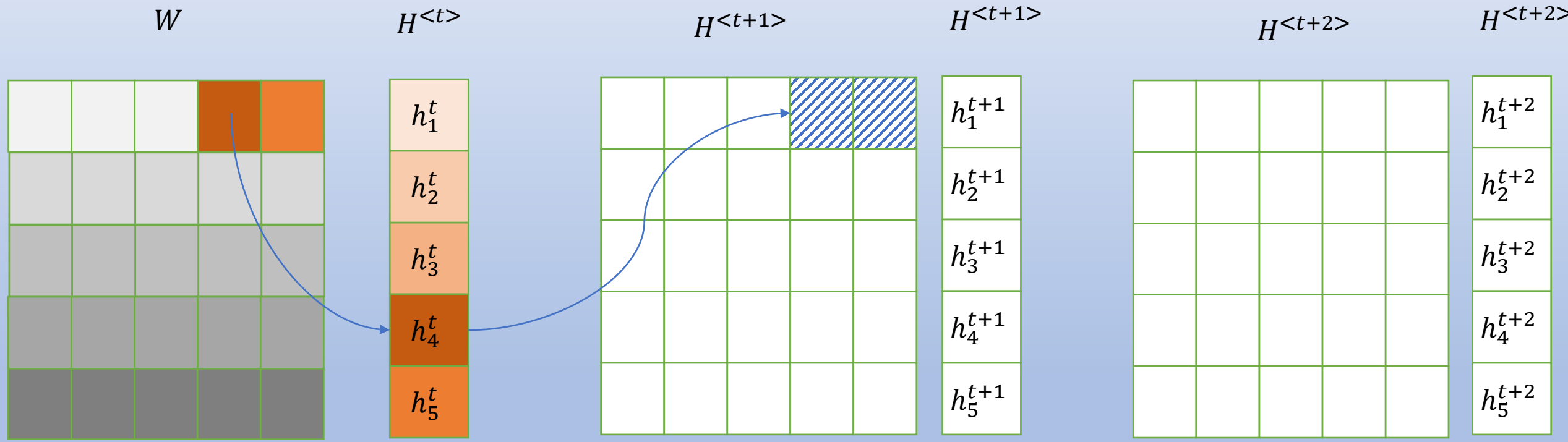
Proposed Approach: Weight Matrix Data Reuse Over Time



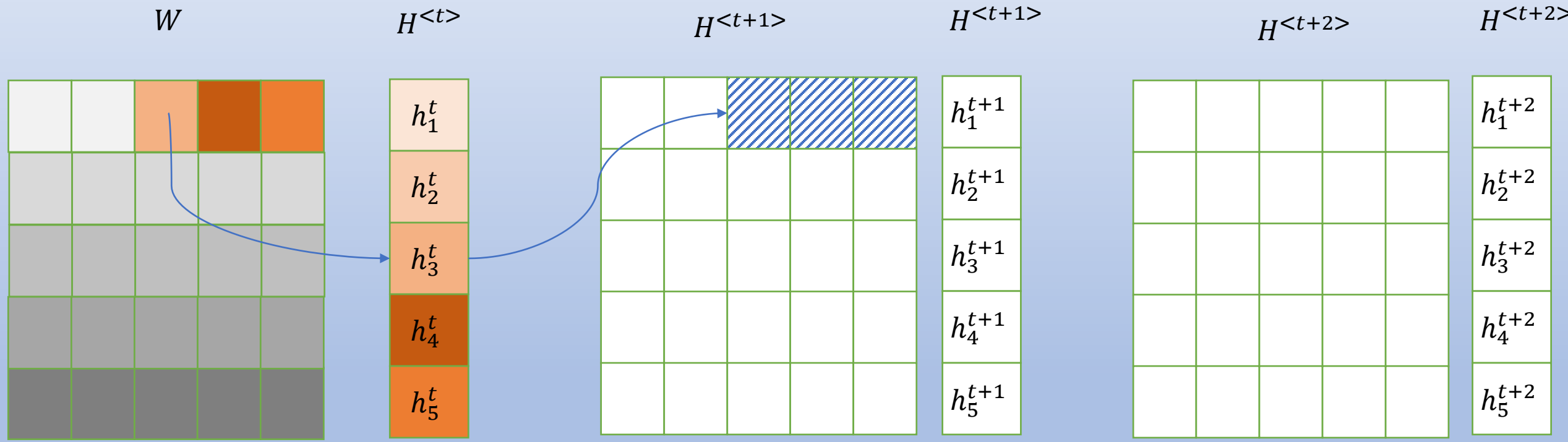
Proposed Approach: Weight Matrix Data Reuse Over Time



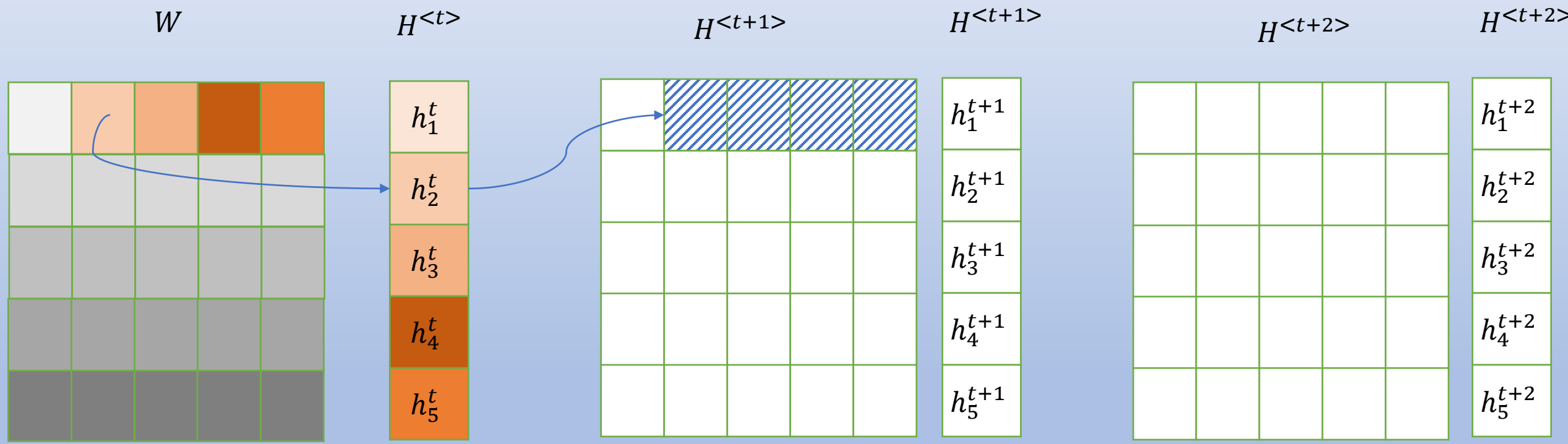
Proposed Approach: Weight Matrix Data Reuse Over Time



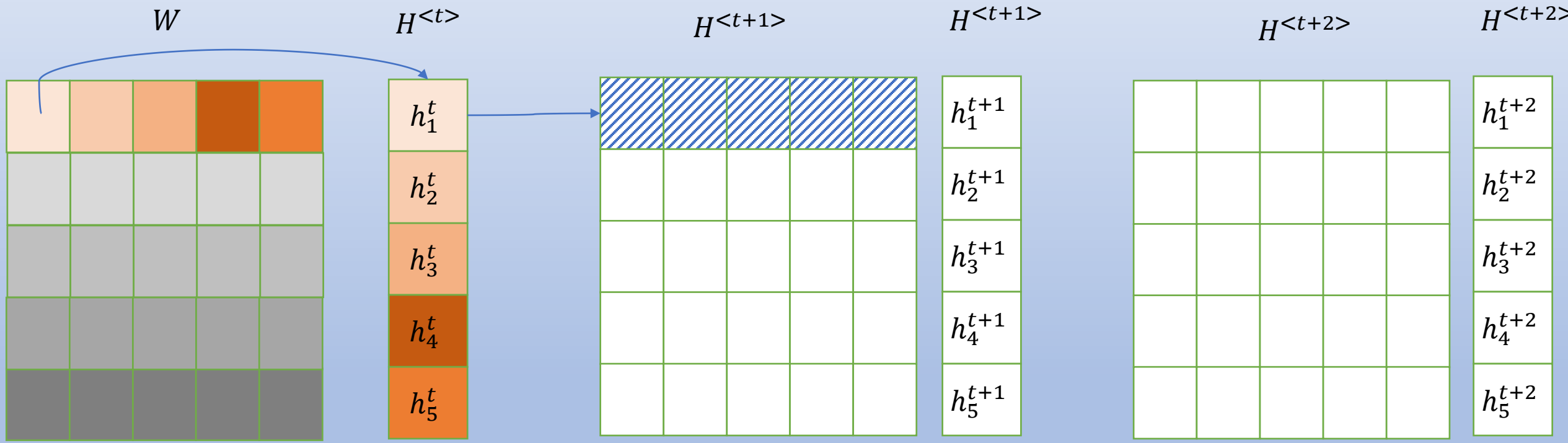
Proposed Approach: Weight Matrix Data Reuse Over Time



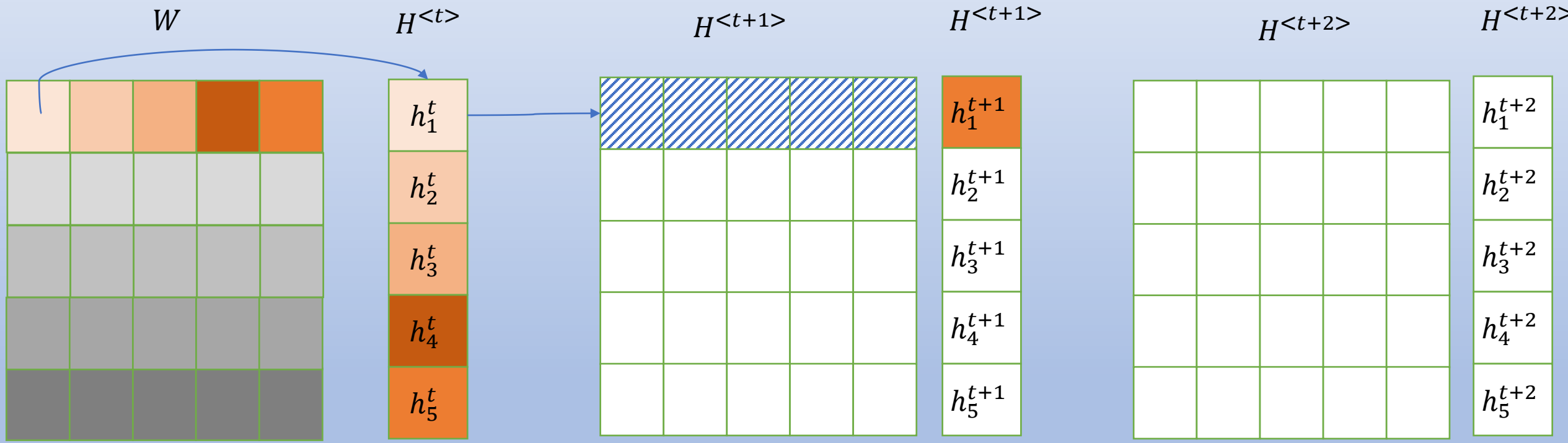
Proposed Approach: Weight Matrix Data Reuse Over Time



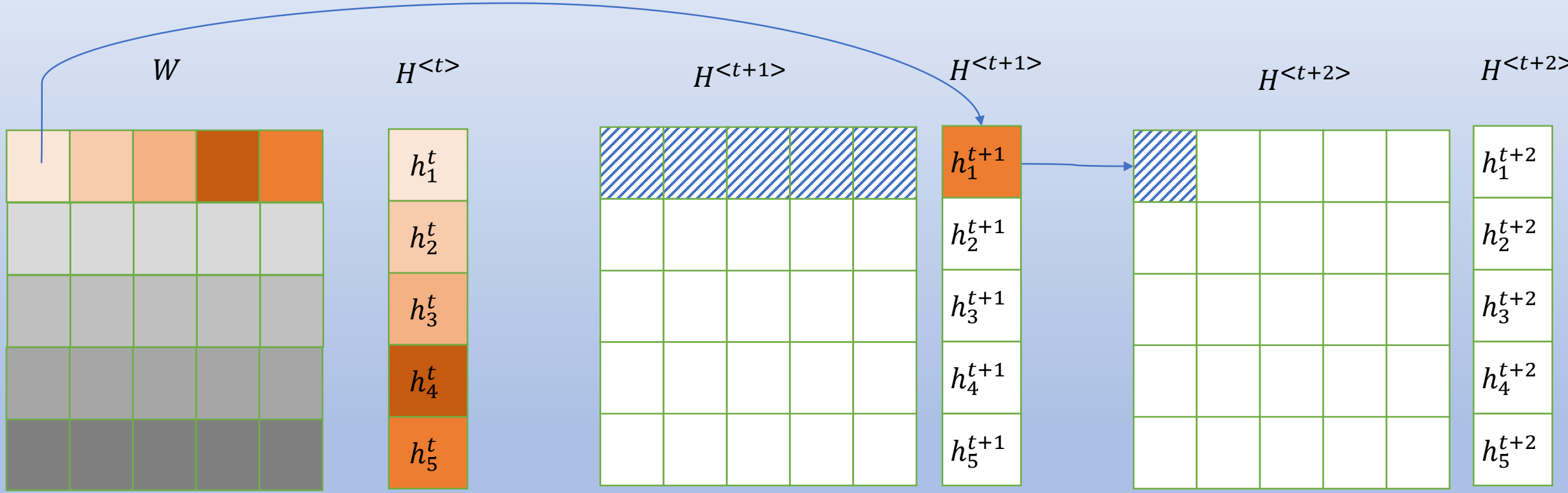
Proposed Approach: Weight Matrix Data Reuse Over Time



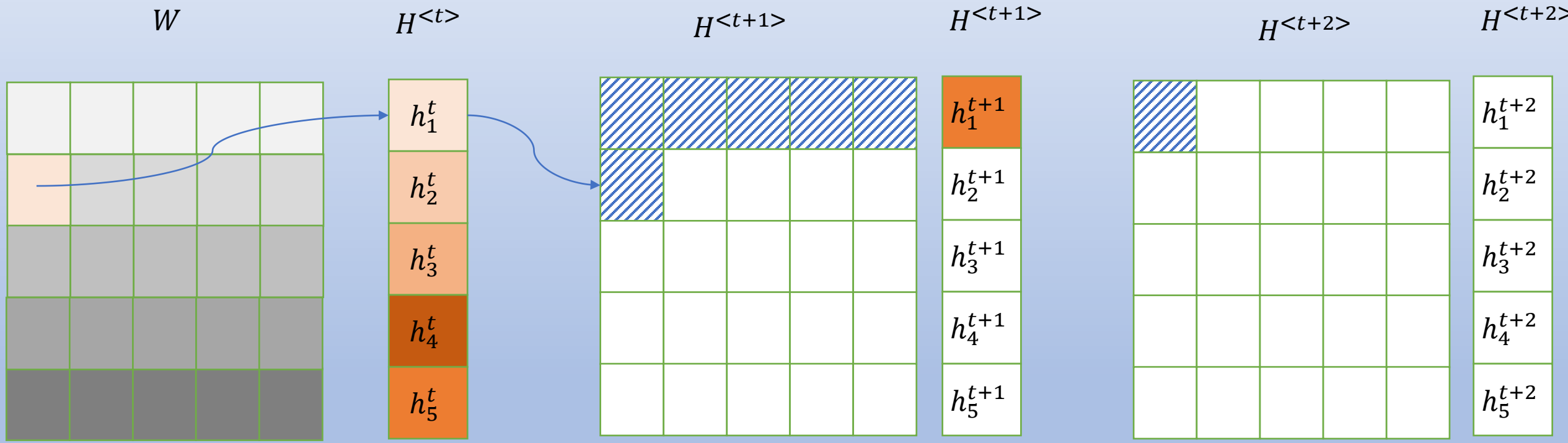
Proposed Approach: Weight Matrix Data Reuse Over Time



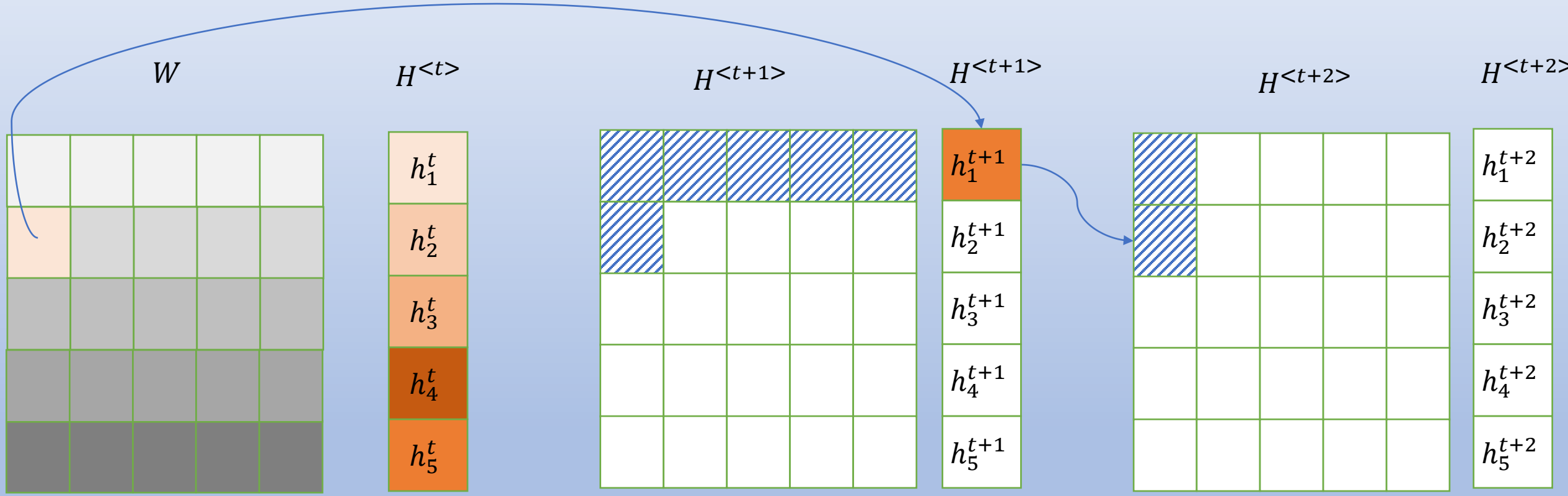
Proposed Approach: Weight Matrix Data Reuse Over Time



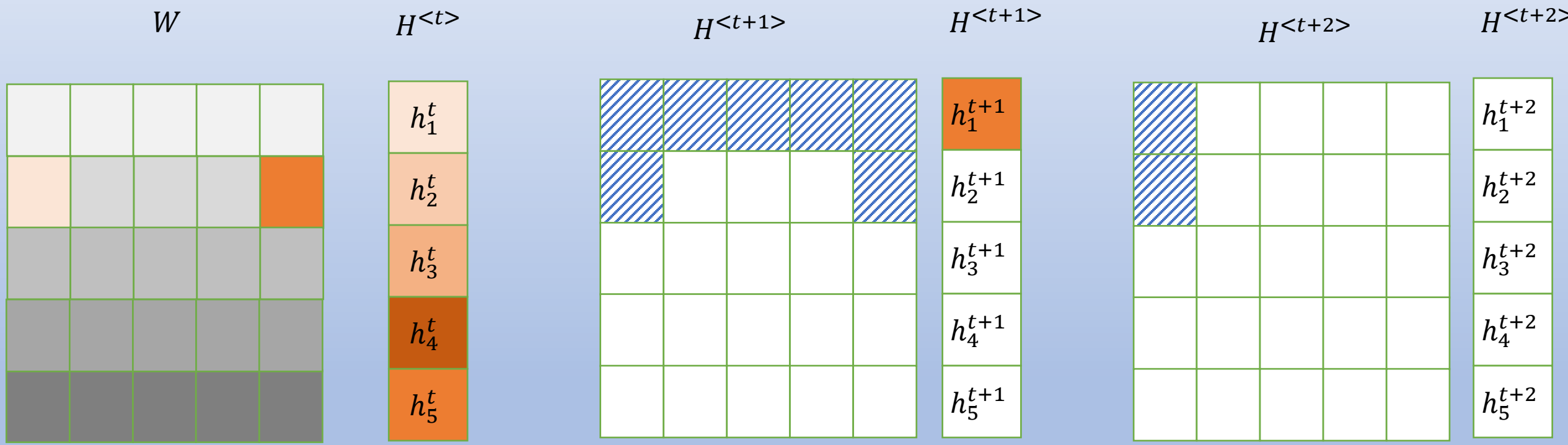
Proposed Approach: Weight Matrix Data Reuse Over Time



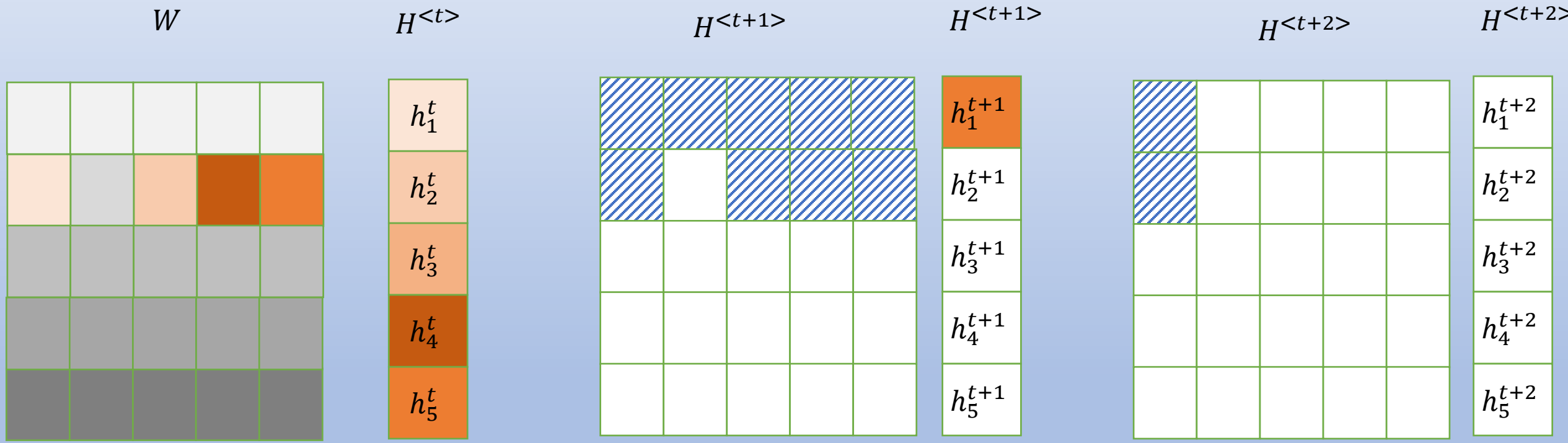
Proposed Approach: Weight Matrix Data Reuse Over Time



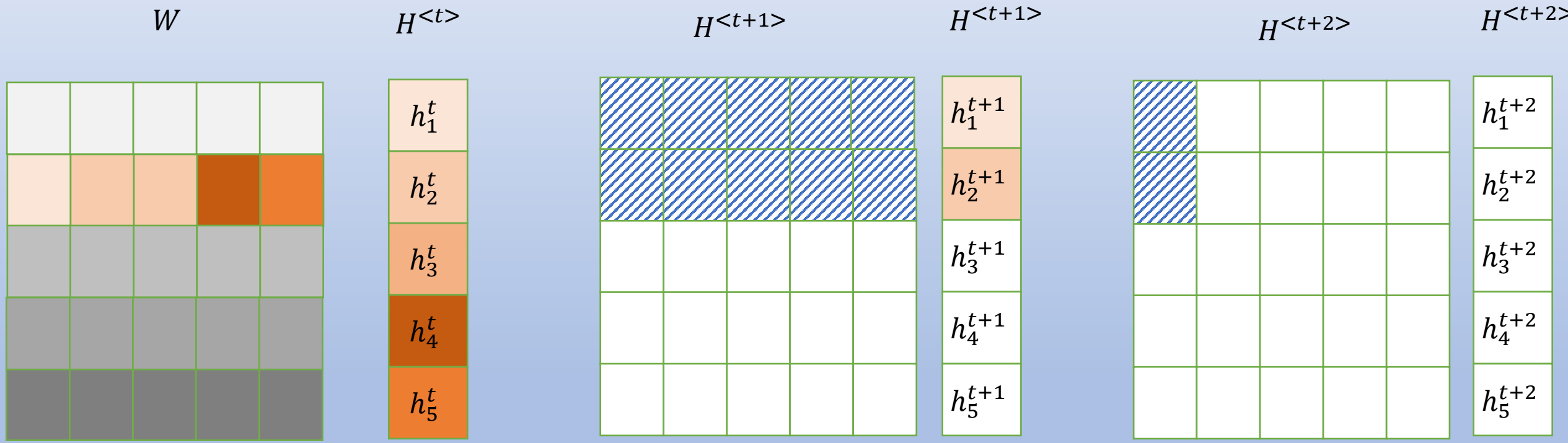
Proposed Approach: Weight Matrix Data Reuse Over Time



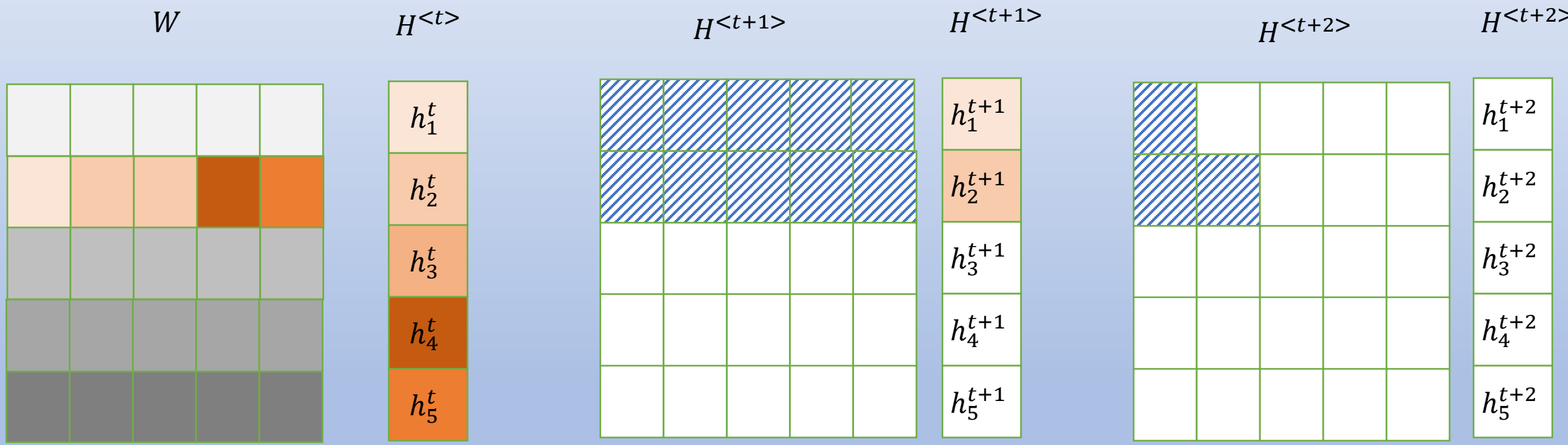
Proposed Approach: Weight Matrix Data Reuse Over Time



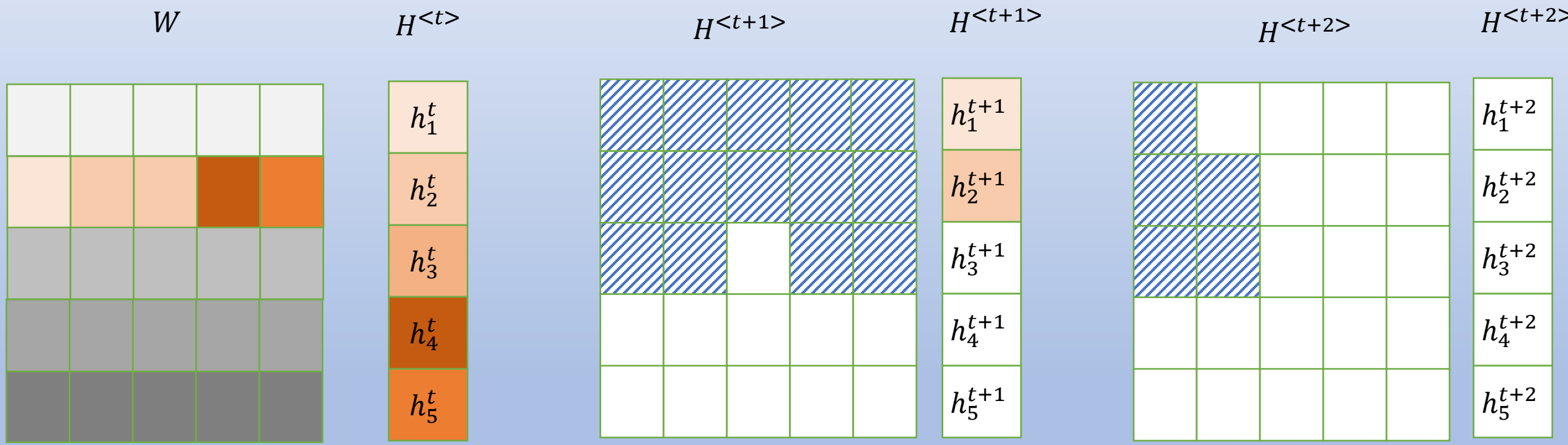
Proposed Approach: Weight Matrix Data Reuse Over Time



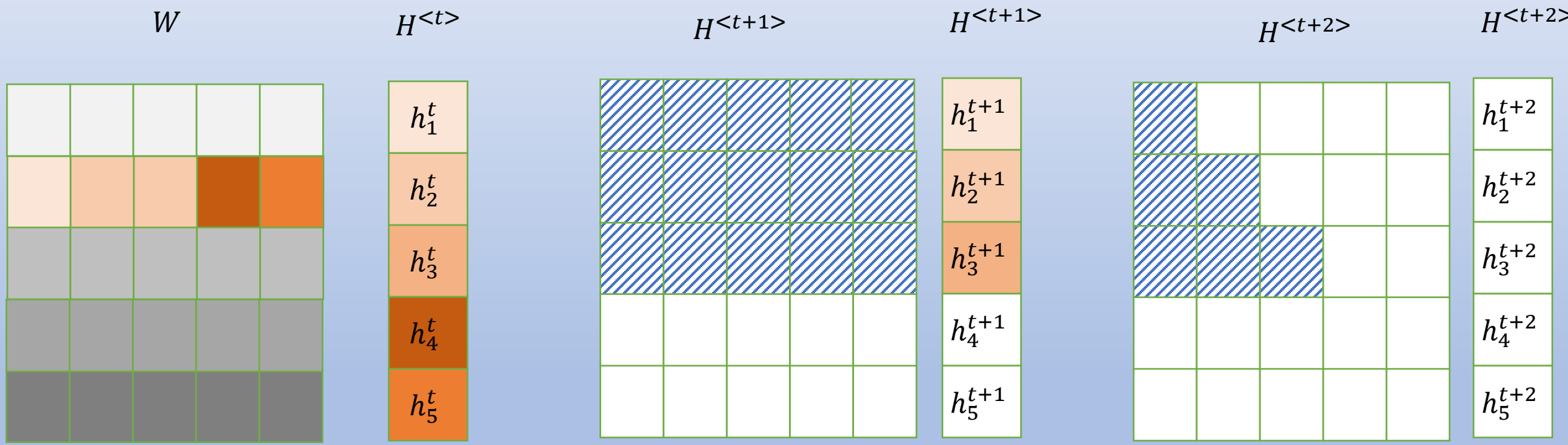
Proposed Approach: Weight Matrix Data Reuse Over Time



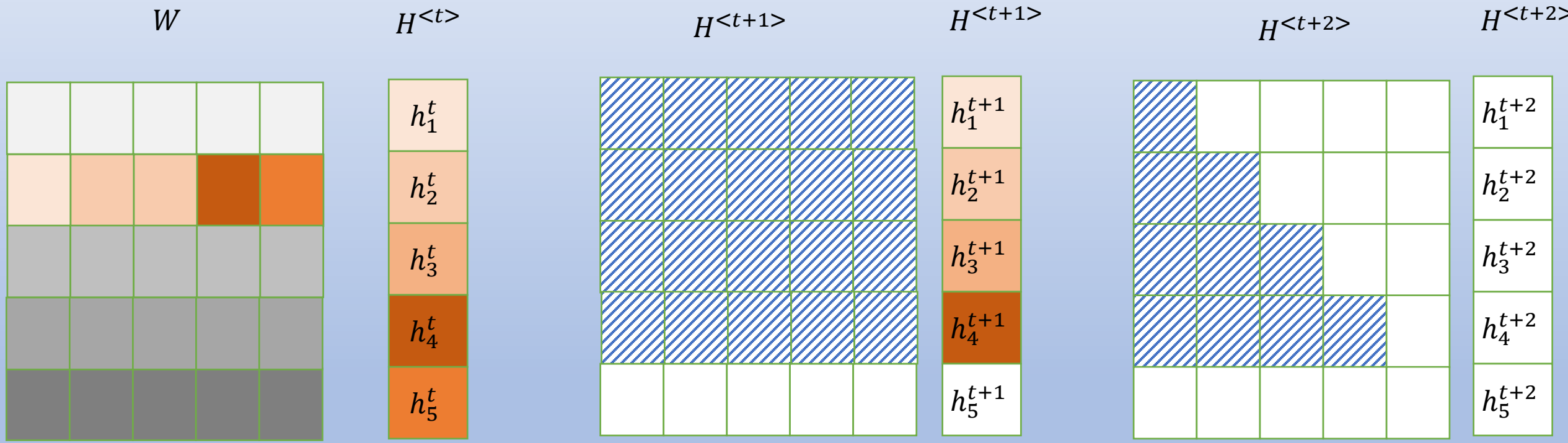
Proposed Approach: Weight Matrix Data Reuse Over Time



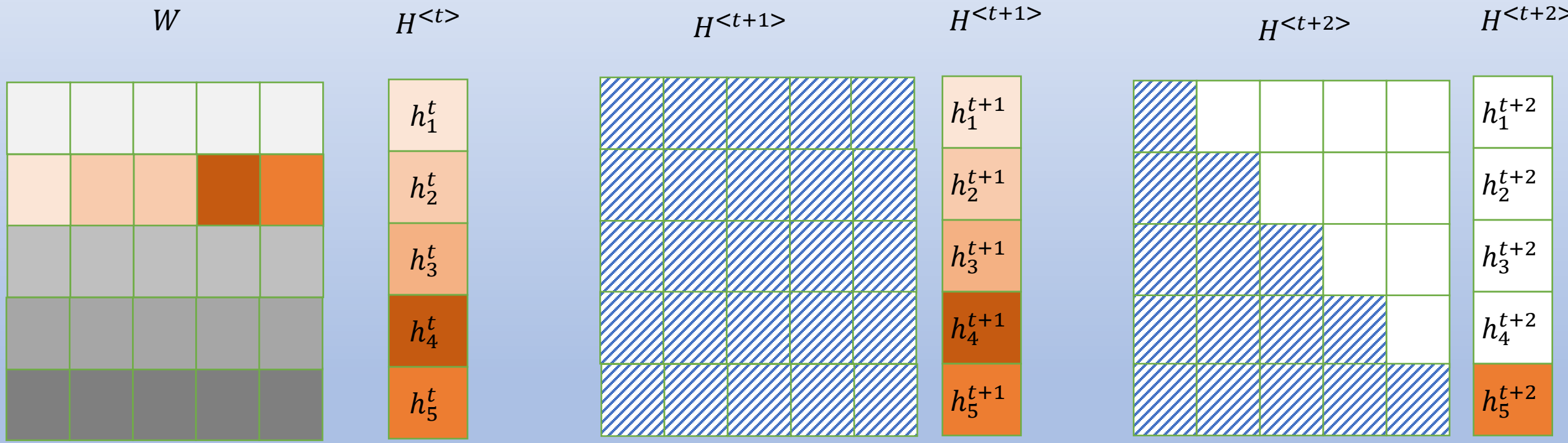
Proposed Approach: Weight Matrix Data Reuse Over Time



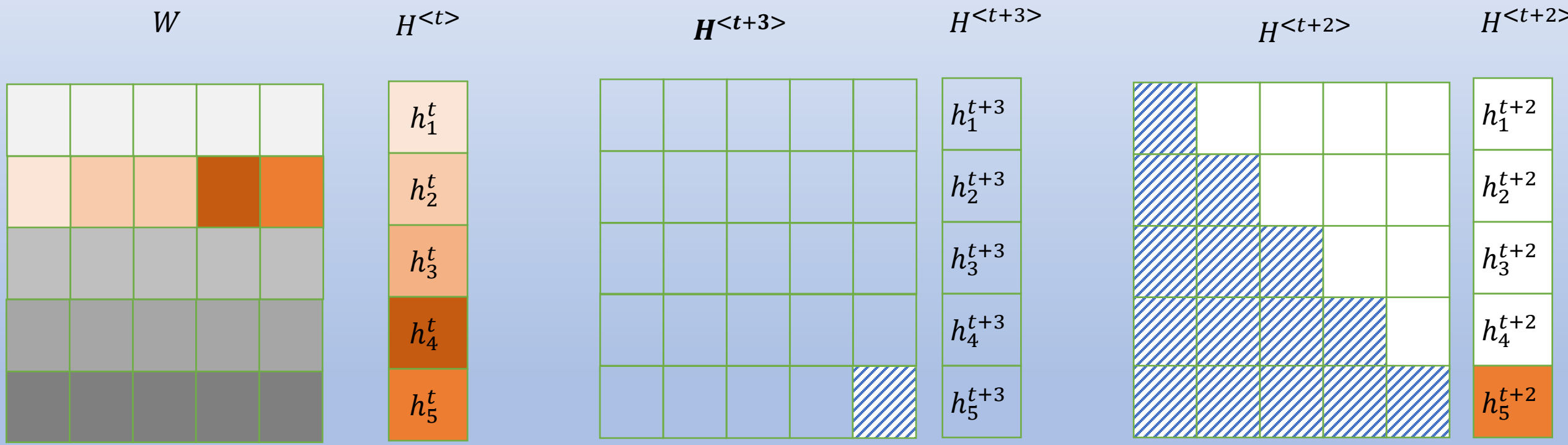
Proposed Approach: Weight Matrix Data Reuse Over Time



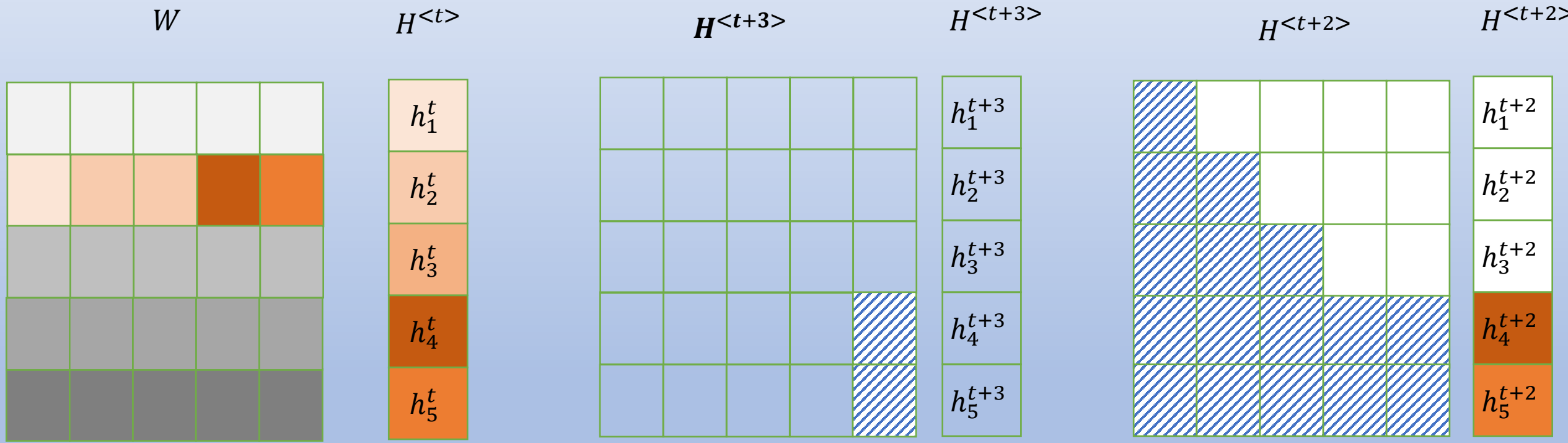
Proposed Approach: Weight Matrix Data Reuse Over Time



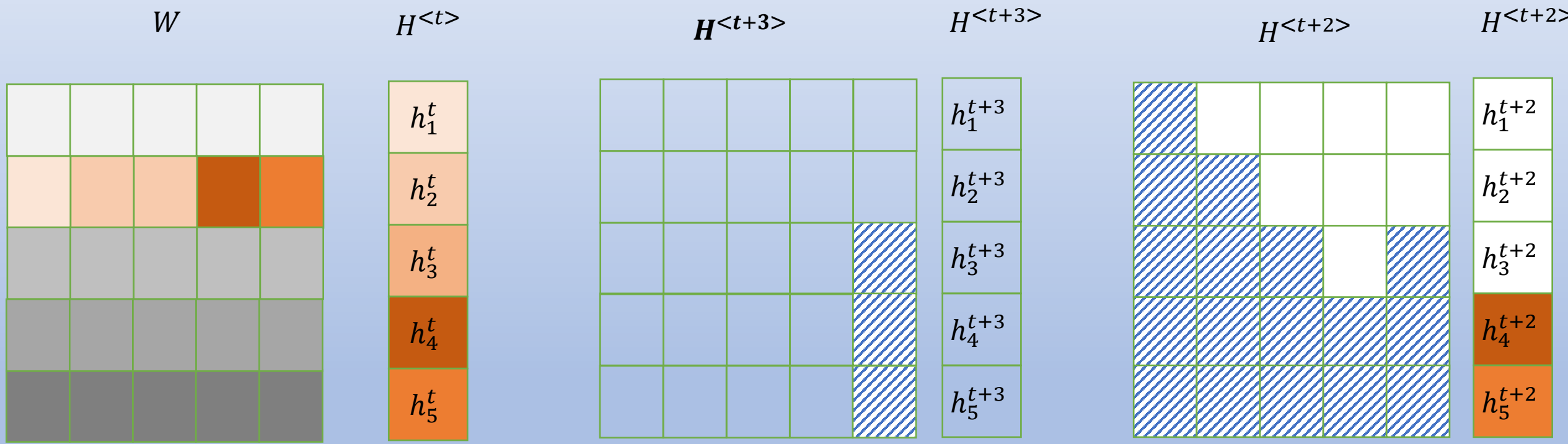
Proposed Approach: Weight Matrix Data Reuse Over Time



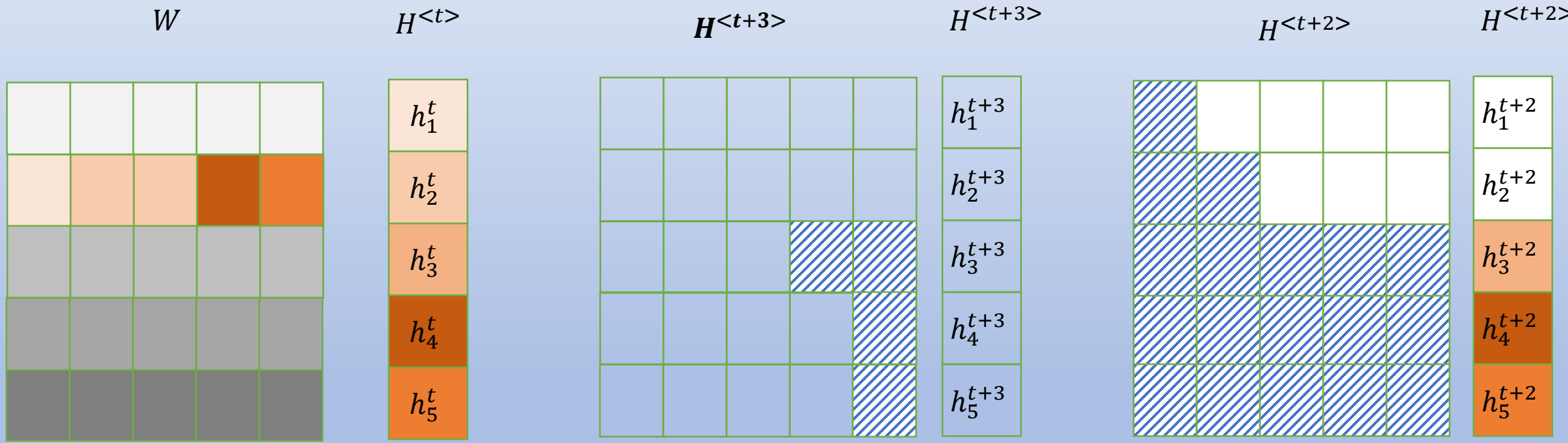
Proposed Approach: Weight Matrix Data Reuse Over Time



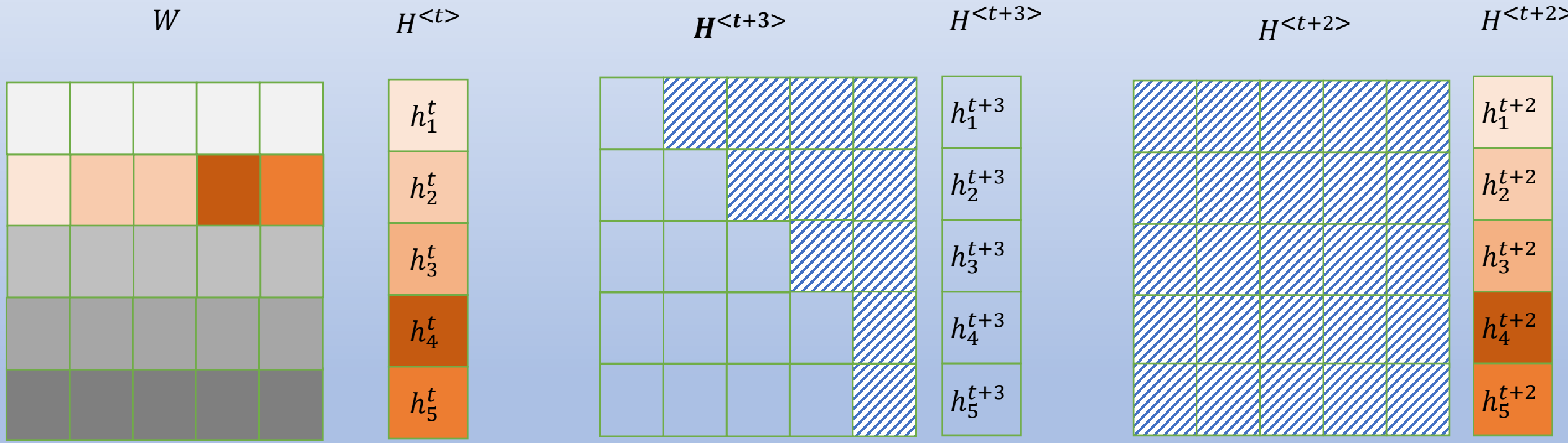
Proposed Approach: Weight Matrix Data Reuse Over Time



Proposed Approach: Weight Matrix Data Reuse Over Time



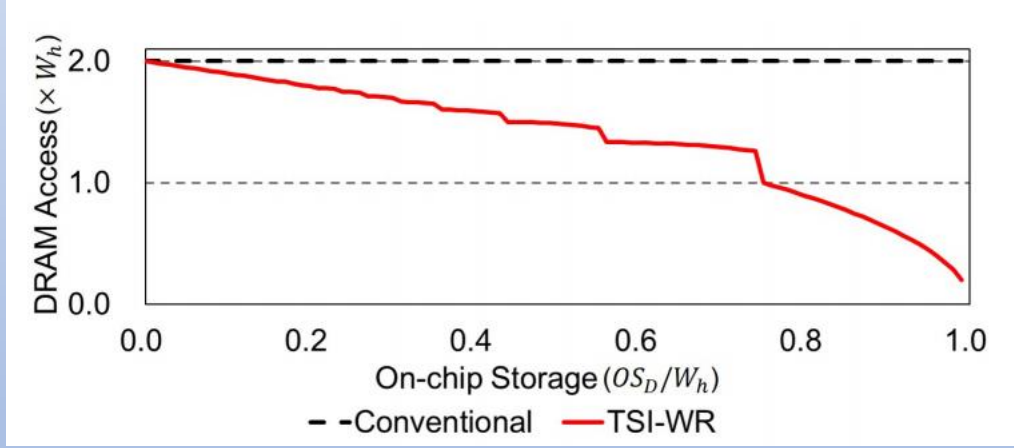
Proposed Approach: Weight Matrix Data Reuse Over Time



Comparison

Previous Work

- Require 79% of Weight Matrix Size storage to get 50% data reuse



- Data reuse depends on on-chip storage size.

Proposed Approach

- With minimal storage (bytes) achieves 50% data reuse.

- Data reuse is independent of on-chip storage size.
- If storage is available then approach is applied to remaining (W_h-M) size for reuse.