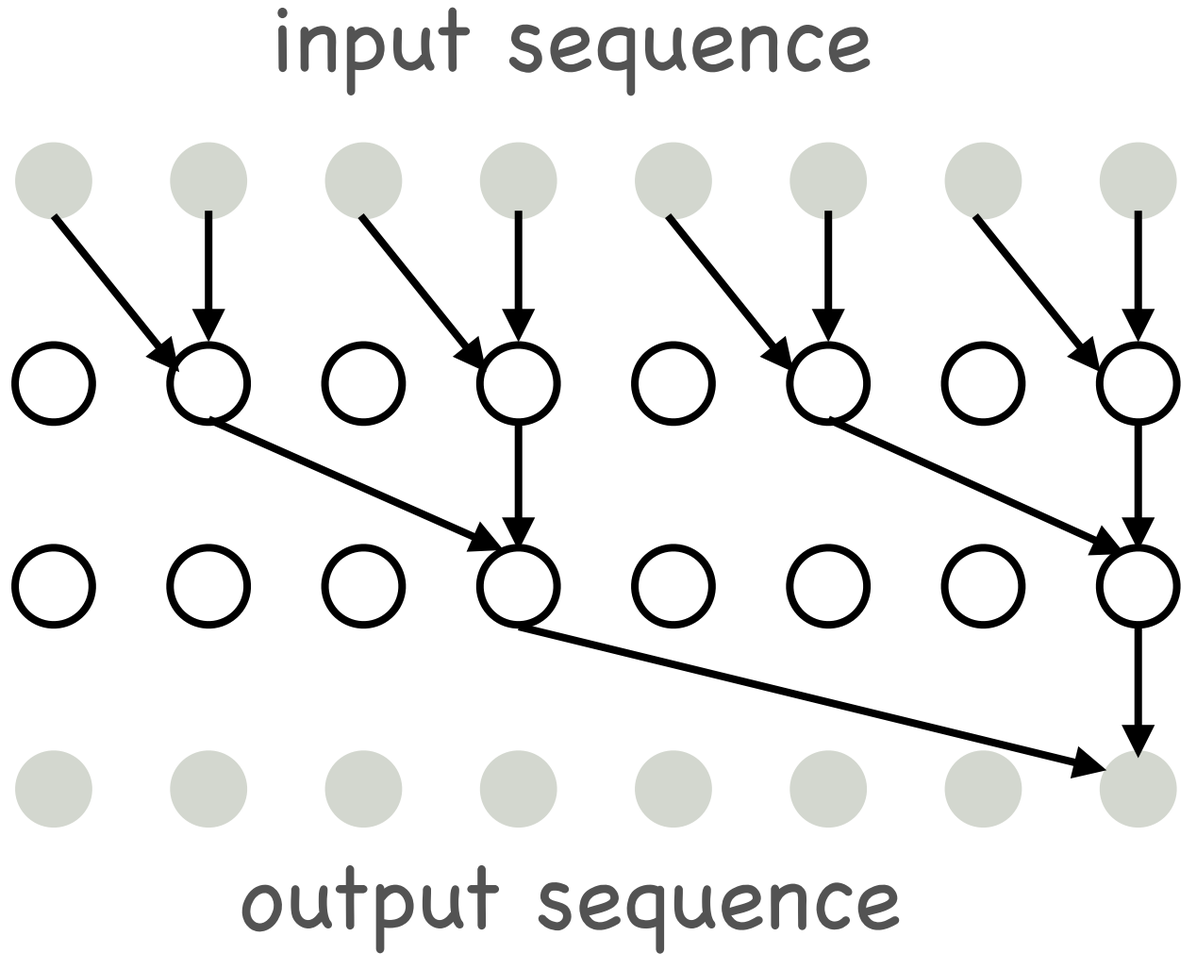
Case study: WaveNet

WaveNet

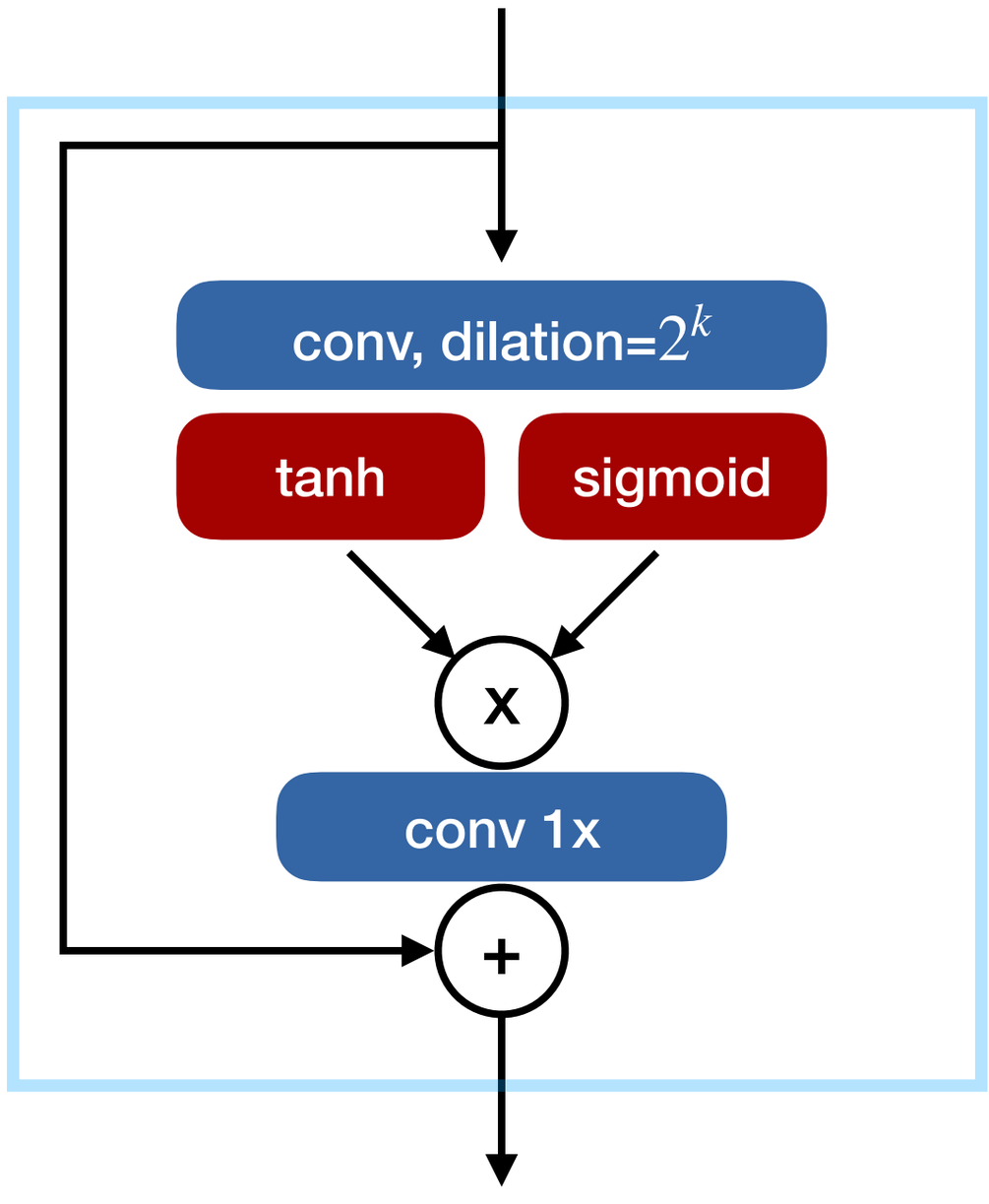
- Autoregressive model for sound synthesis and speech recognition
  - Generates raw waveform
    - Quantized in 8-bit
  - $P(y_t \mid x, y_0, \ldots, y_{t-1})$



WaveNet: A generative model for raw audio, Van Den Oord et al., *arXiv* 2016

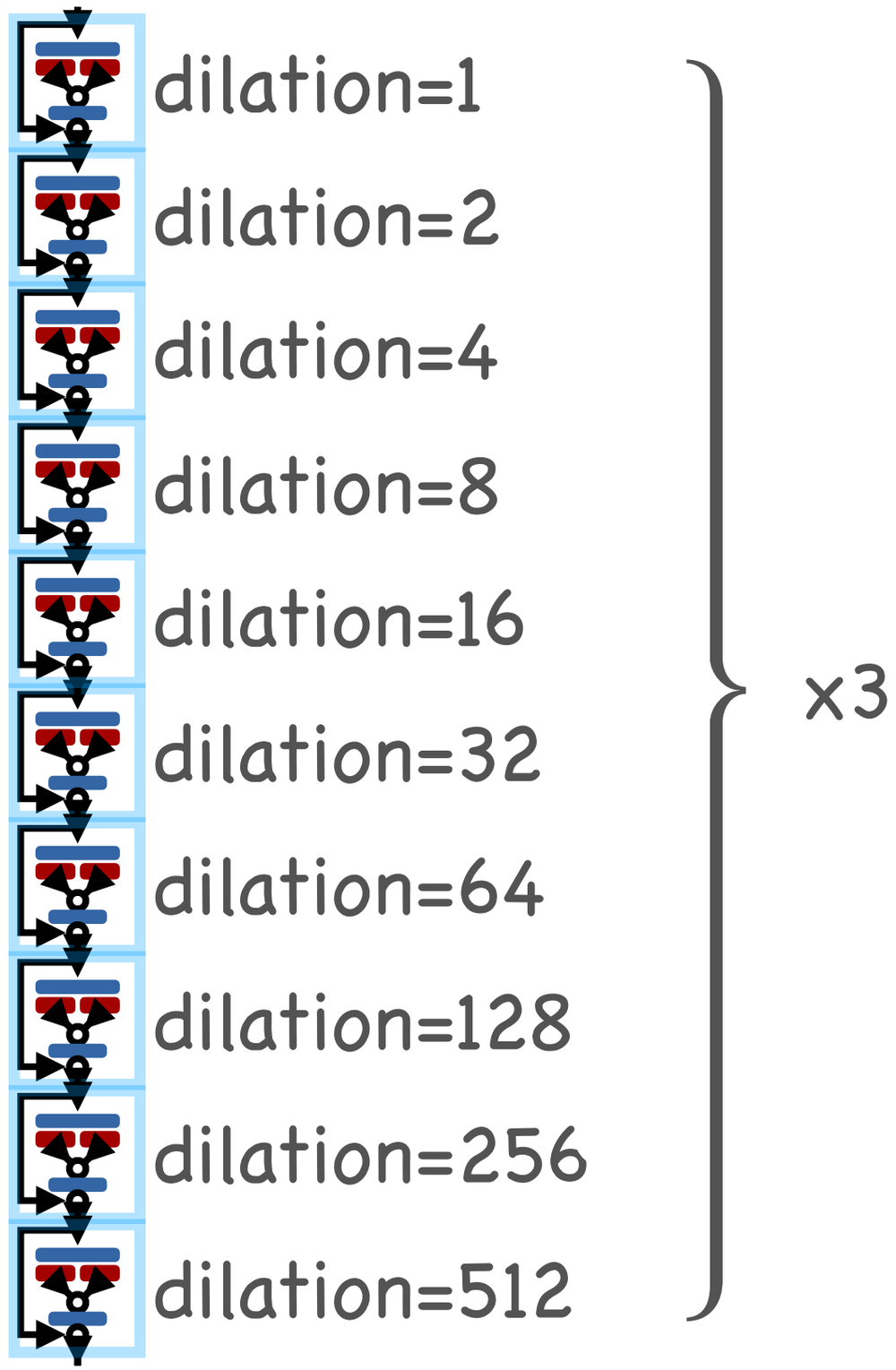
WaveNet - basic building block

- Dilated causal convolution
- Gated activation units



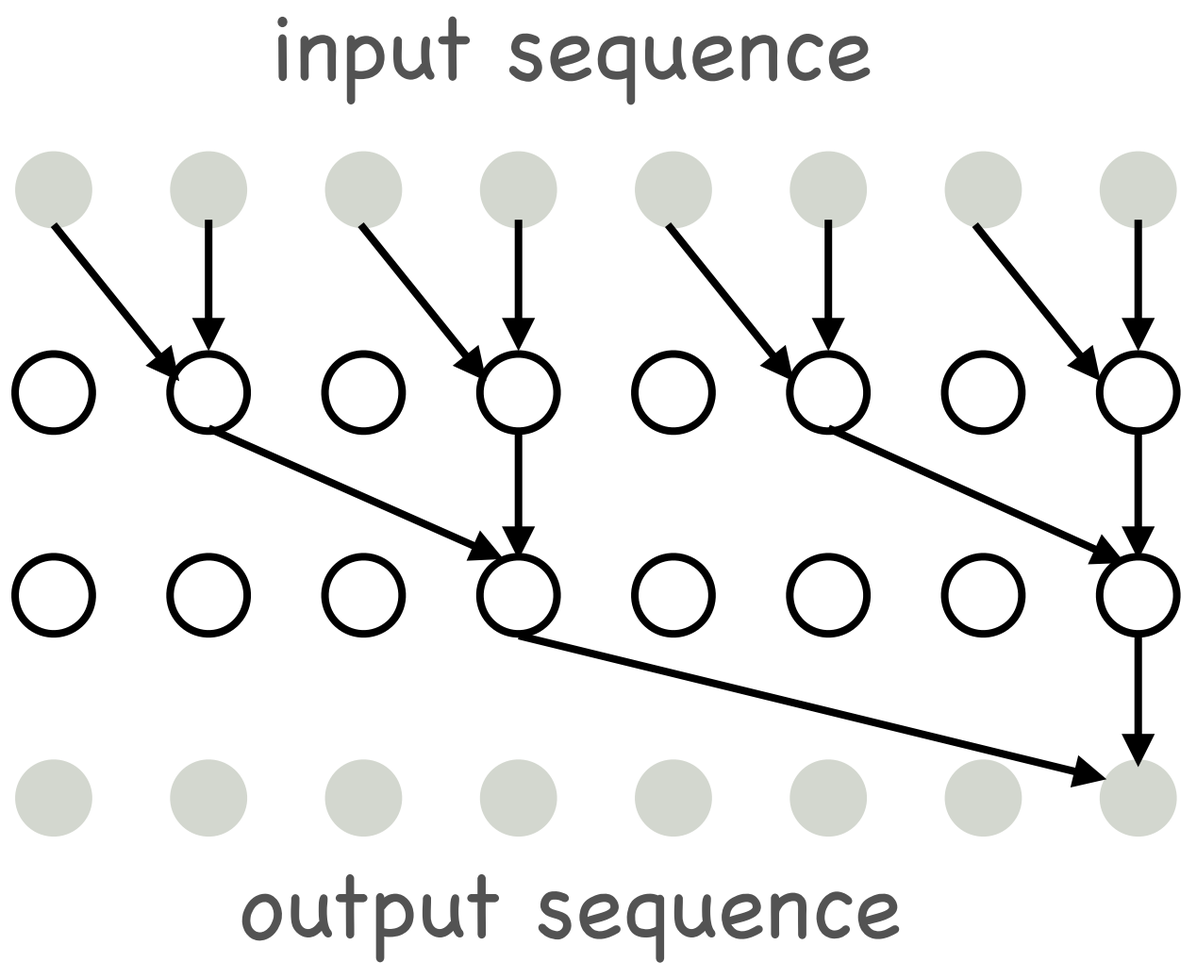
WaveNet

- Input
  - Causal generation y
- Output
  - $P(y_t \mid x, y_0, \ldots, y_{t-1})$



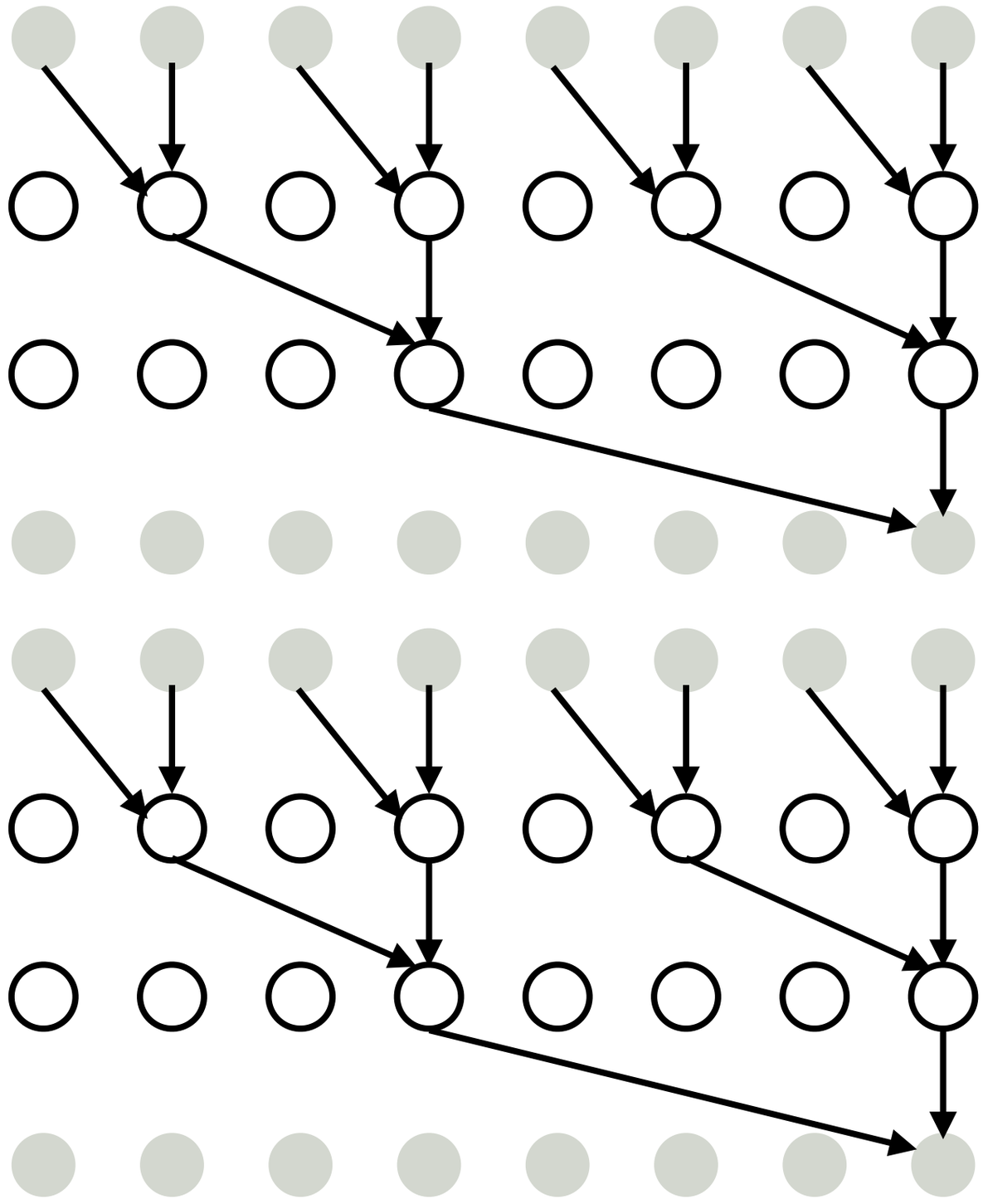
WaveNet

- State-of-the-art music and English speech generation
- Slow



Parallel WaveNet

- Inverse Autoregressive Flow (IAF)
  - Transform noise into sound
  - Single feed forward pass
    - No sampling
- Trained to mimic original WaveNet
- 500k samples / sec, 10x real time
  - Used by Google Assistant



Parallel WaveNet: Fast High-Fidelity Speech Synthesis, van den Oord et al., arXiv 2017