EPFL

PARALLEL SYSTEMS ARCHITECTURE LAB

Derin Sinir Ağları için Değişken Boyutlu Sistolik Diziler

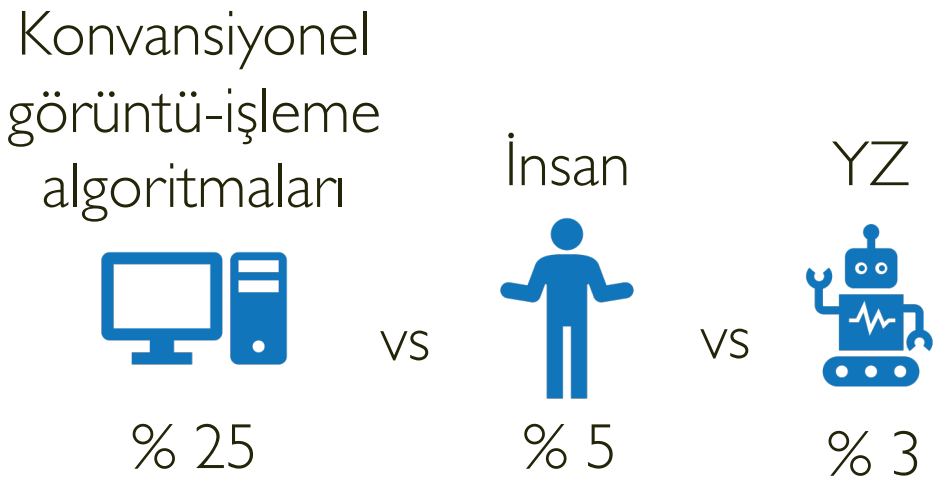
Ahmet Caner Yüzügüler – Doktora adayı

Prof. Babak Falsafi - Danışman

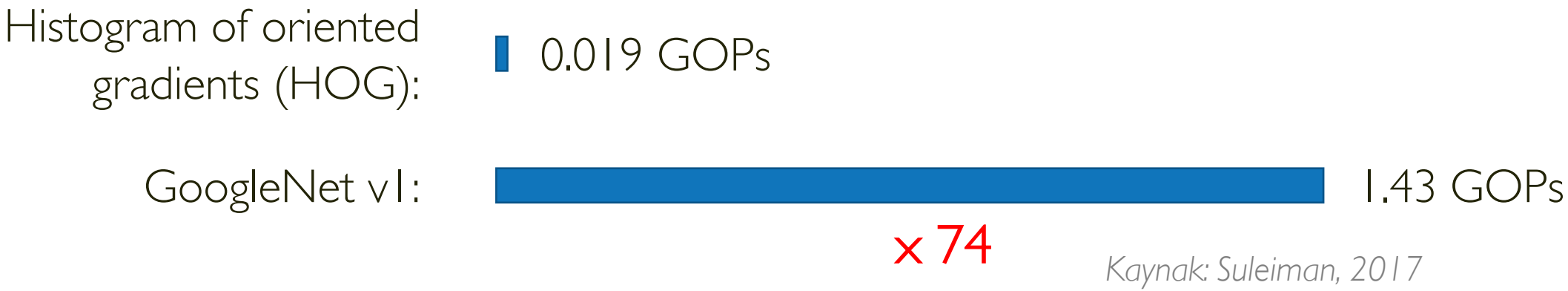
Derin Sinir Ağları (DSA)

PARALLEL SYSTEMS ARCHITECTURE LAB

- ✓ Hızlı algoritma tasarımı
  - Otomatik öznitelik (feature) cikarimi
- ✓ İnsan-üstü isabet oranı
  - Imagenet veriseti hata oranı:
- ✗ Çok yüksek bilgisayım gereksinimi



Kaynak: Sze, 2017

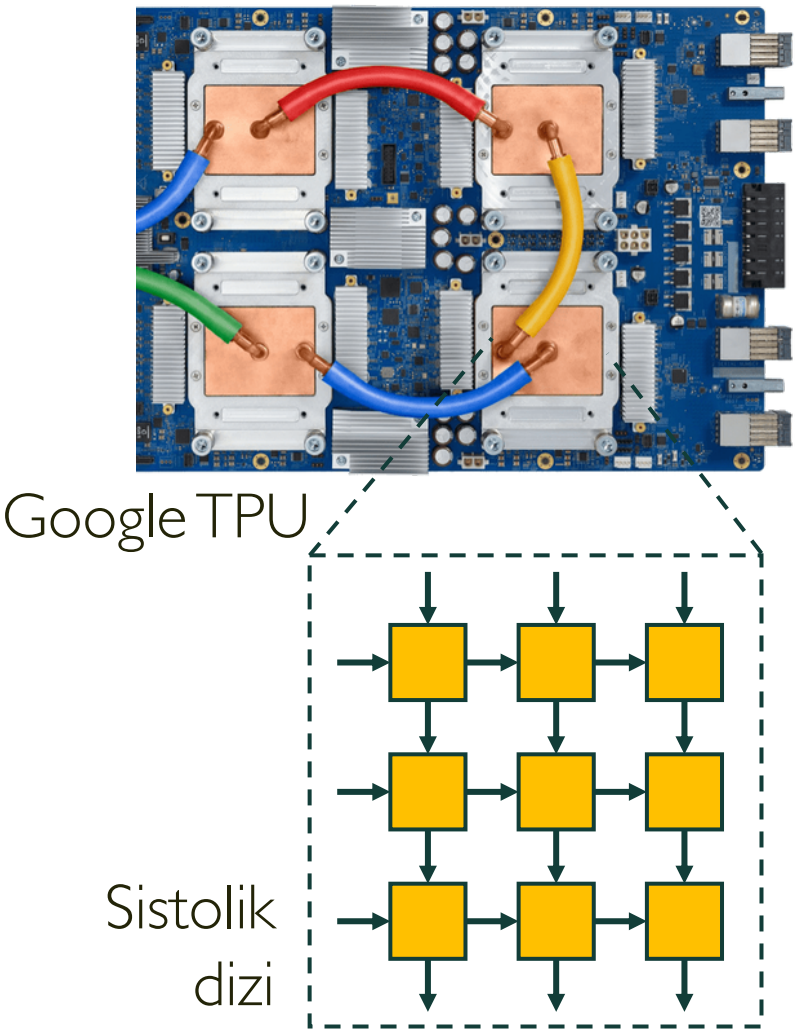


Kaynak: Suleiman, 2017

Sistolik Diziler

- Google'ın DSA'lar için ASIC çözümü:
  - Tensor Processing Unit (TPU)
  - 256 x 256 sistolik dizi
- Neden sistolik dizi?
  - ✓ Yüksek verimli işlem kapasitesi
    - 92 TeraOps/s @ 40 W
  - ✓ Ölçeklenebilirlik
  - ✓ Kolay tasarım – hızlı geliştirme süresi
- Dezavantajı
  - ✗ Veri paylaşımında düşük esneklik
  - ✗ Düşük kapasite kullanım oranı
    - < %30 (GoogleNet)

Resim: https://cloud.google.com/tpu/



Önerdiğimiz Çözüm

- İri parçalı yeniden yapılandırılabilir mimari (Coarse-grained reconfigurable architecture)

Bloklar arasında programlanabilir bağlantılar

↓

Esnek veri aktarım altyapısı

↓

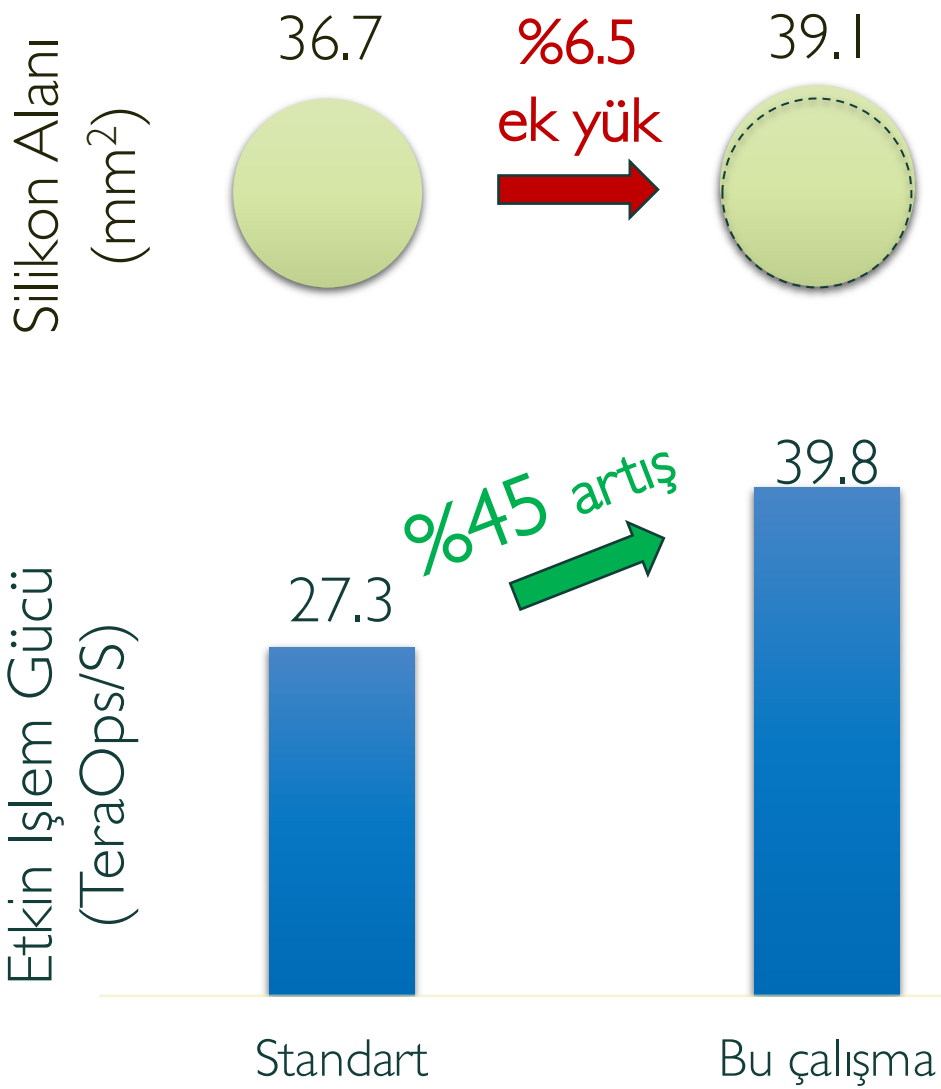
Farklı boyutlardaki DSA katmanlarına daha iyi uyum

↓

Yüksek kapasite kullanımı

↓

Yüksek işlem gücü



İçerik

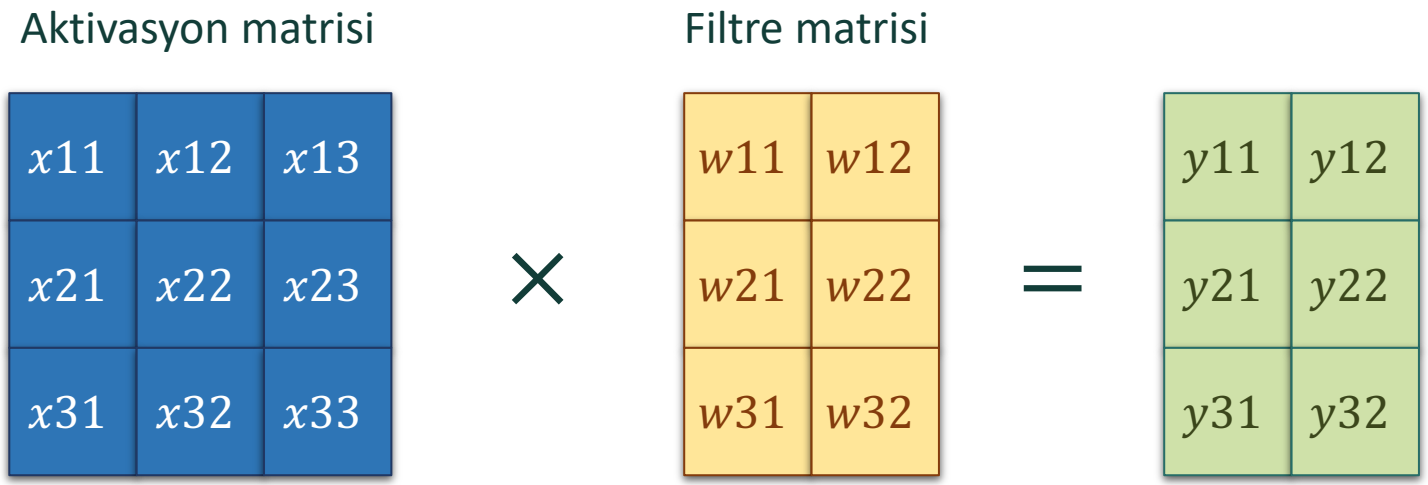
PARALLEL SYSTEMS ARCHITECTURE LAB

- Giriş
- Sistolik diziler
- Düşük kapasite kullanım oranı sebebleri
- Önerilen çözüm
- Sonuçlar

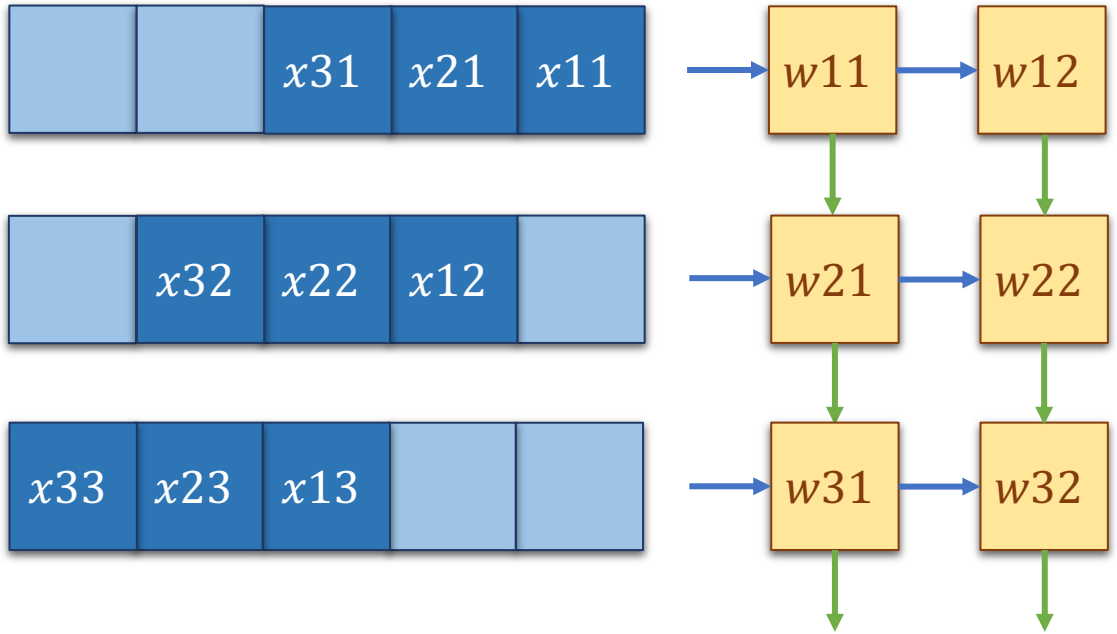
Sistolik Diziler – Çalışma Prensibi

- Temelde gerçekleştirdikleri işlem: **matris-matris çarpımı**
- DSA'larını oluşturan işlemlerin büyük çoğunluğu (>%95):
  - **Aktivasyon** ve **filtre** matrislerinin çarpımı

- Örnek:



Sistolik Diziler – Çalışma Prensibi



Aktivasyon matrisi

$x11$	$x12$	$x13$
$x21$	$x22$	$x23$
$x31$	$x32$	$x33$

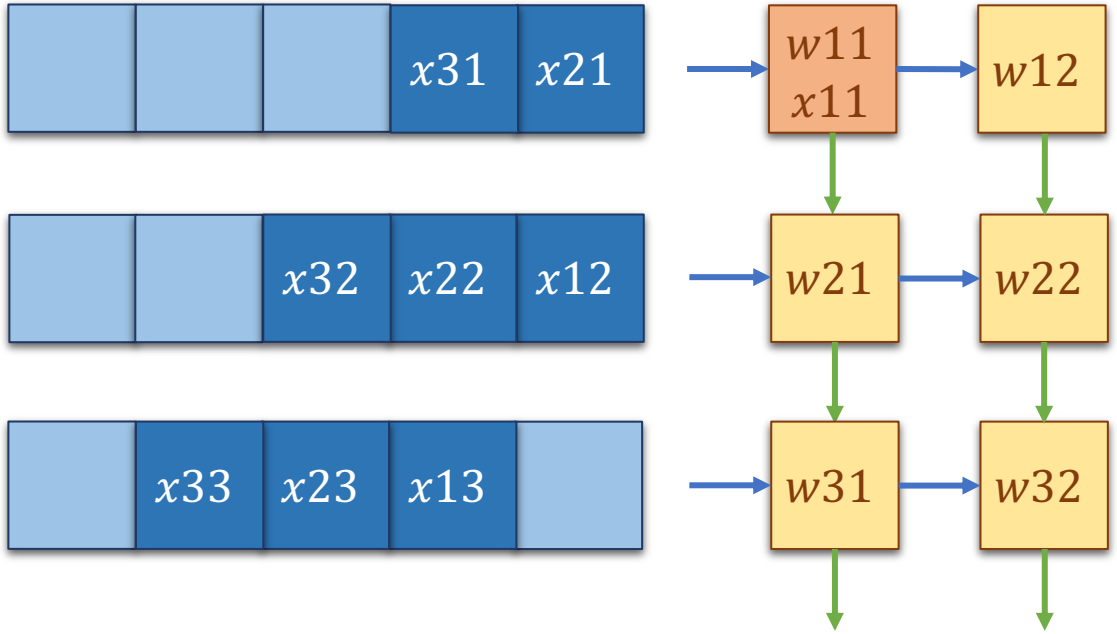
$\times$

Filtre matrisi

$w11$	$w12$
$w21$	$w22$
$w31$	$w32$

$=$

Sistolik Diziler – Çalışma Prensibi



Aktivasyon matrisi

$x11$	$x12$	$x13$
$x21$	$x22$	$x23$
$x31$	$x32$	$x33$

$\times$

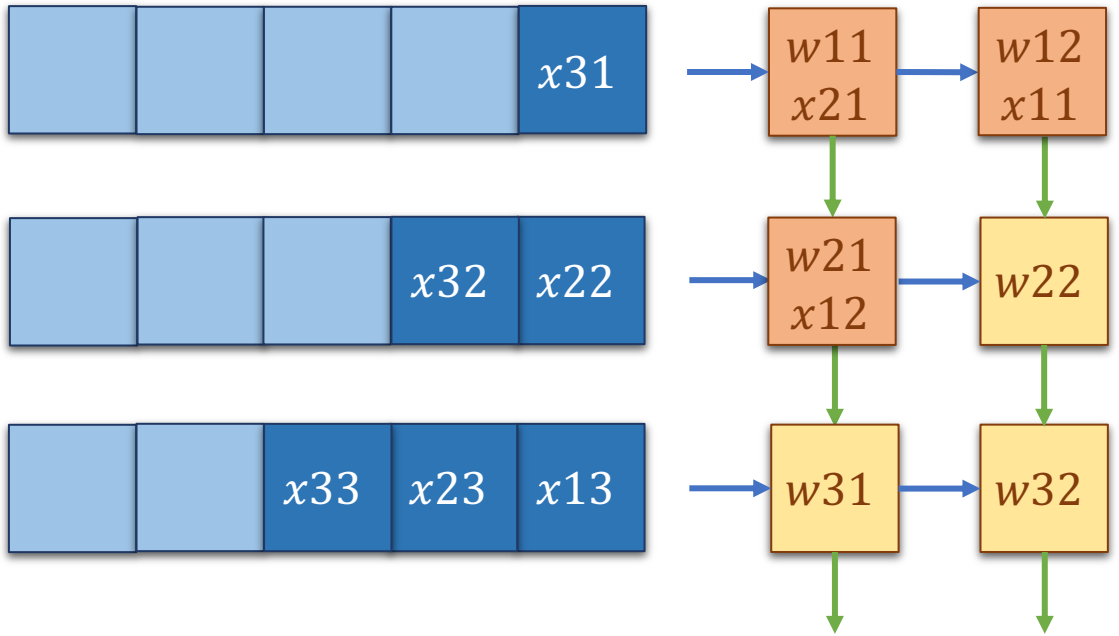
Filtre matrisi

$w11$	$w12$
$w21$	$w22$
$w31$	$w32$

$=$

$$y11 = w11\,x11$$

Sistolik Diziler – Çalışma Prensibi



Aktivasyon matrisi

$x11$	$x12$	$x13$
$x21$	$x22$	$x23$
$x31$	$x32$	$x33$

$\times$

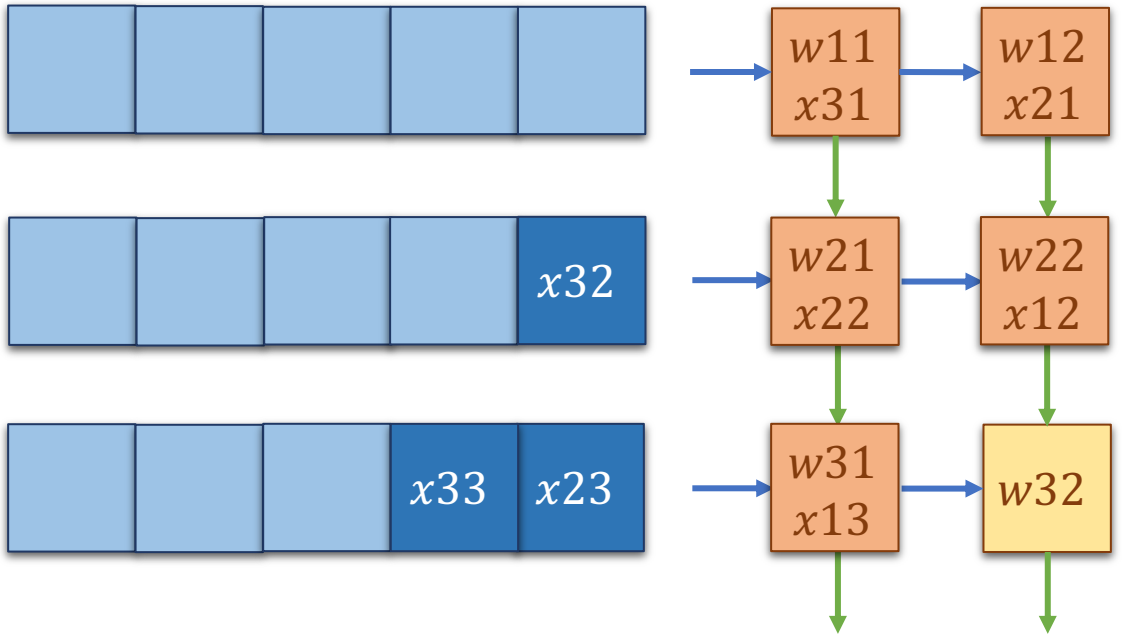
Filtre matrisi

$w11$	$w12$
$w21$	$w22$
$w31$	$w32$

$=$

$$\mathbf{y11} = w11\,x11 + w21\,x12$$
$$\mathbf{y12} = w12\,x11$$
$$\mathbf{y21} = w11\,x21$$

Sistolik Diziler – Çalışma Prensibi



Aktivasyon matrisi

$x11$	$x12$	$x13$
$x21$	$x22$	$x23$
$x31$	$x32$	$x33$

$\times$

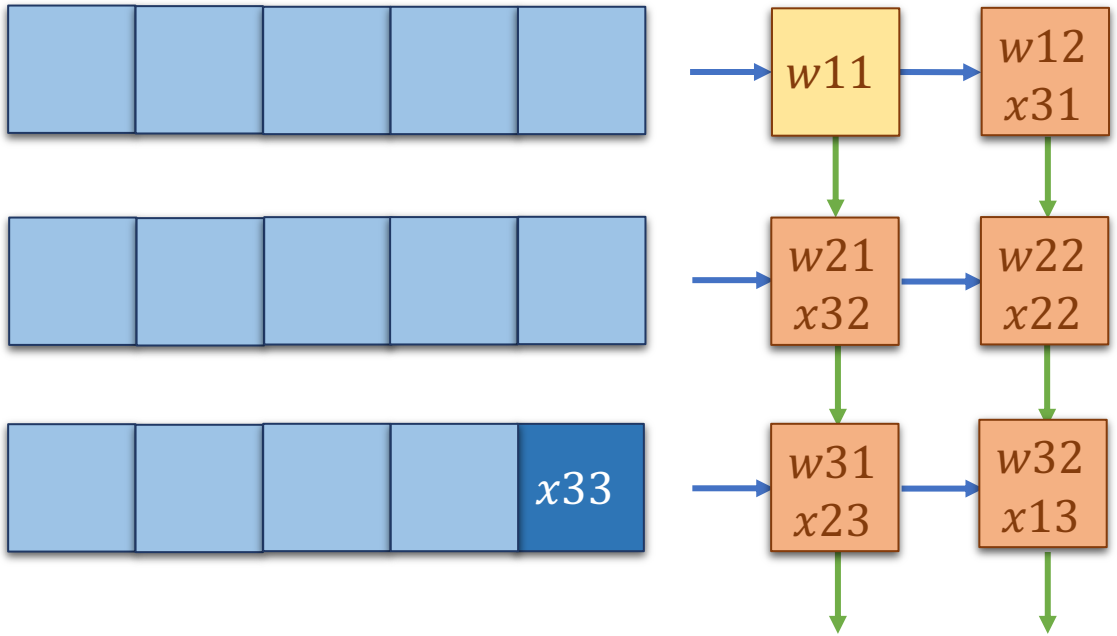
Filtre matrisi

$w11$	$w12$
$w21$	$w22$
$w31$	$w32$

$=$

$$y11 = w11\,x11 + w21\,x12 + w31\,x13$$
$$y12 = w12\,x11 + w22\,x12$$
$$y21 = w11\,x21 + w21\,x22$$
$$y22 = w12\,x21$$
$$y31 = w11\,x31$$

Sistolik Diziler – Çalışma Prensibi



Aktivasyon matrisi

$x11$	$x12$	$x13$
$x21$	$x22$	$x23$
$x31$	$x32$	$x33$

$\times$

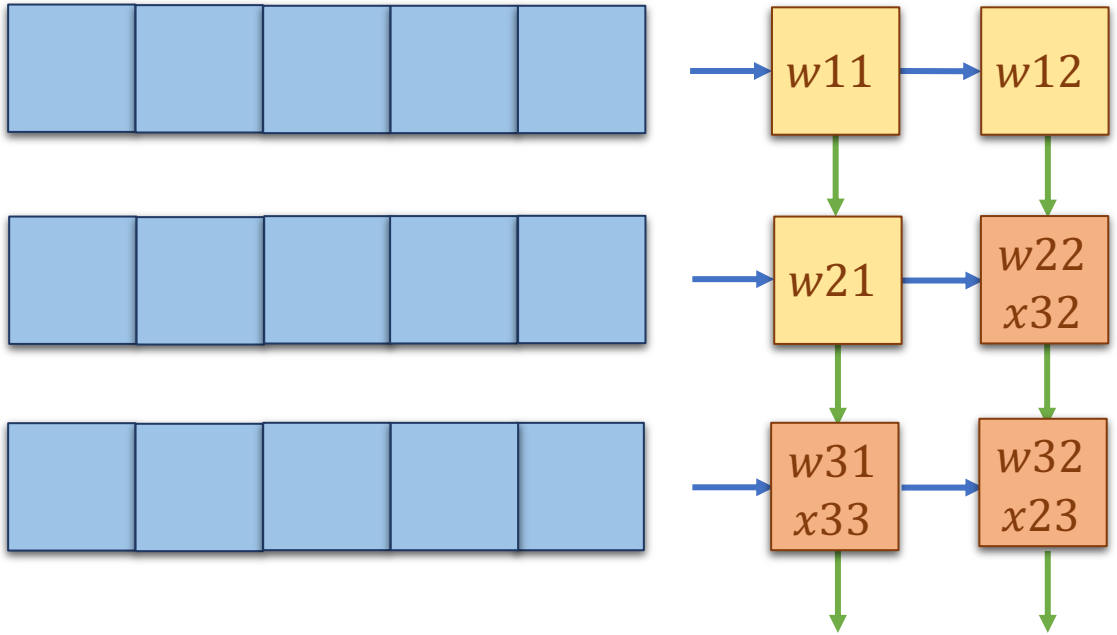
Filtre matrisi

$w11$	$w12$
$w21$	$w22$
$w31$	$w32$

$=$

$$
\begin{aligned}
\boldsymbol{y11} &= w11\,x11 + w21\,x12 + w31\,x13 \\
\boldsymbol{y12} &= w12\,x11 + w22\,x12 + w32\,x13 \\
\boldsymbol{y21} &= w11\,x21 + w21\,x22 + w31\,x23 \\
\boldsymbol{y22} &= w12\,x21 + w22\,x22 \\
\boldsymbol{y31} &= w11\,x31 + w21\,x32 \\
\boldsymbol{y32} &= w12\,x31
\end{aligned}
$$

Sistolik Diziler – Çalışma Prensibi



Aktivasyon matrisi

$x11$	$x12$	$x13$
$x21$	$x22$	$x23$
$x31$	$x32$	$x33$

$\times$

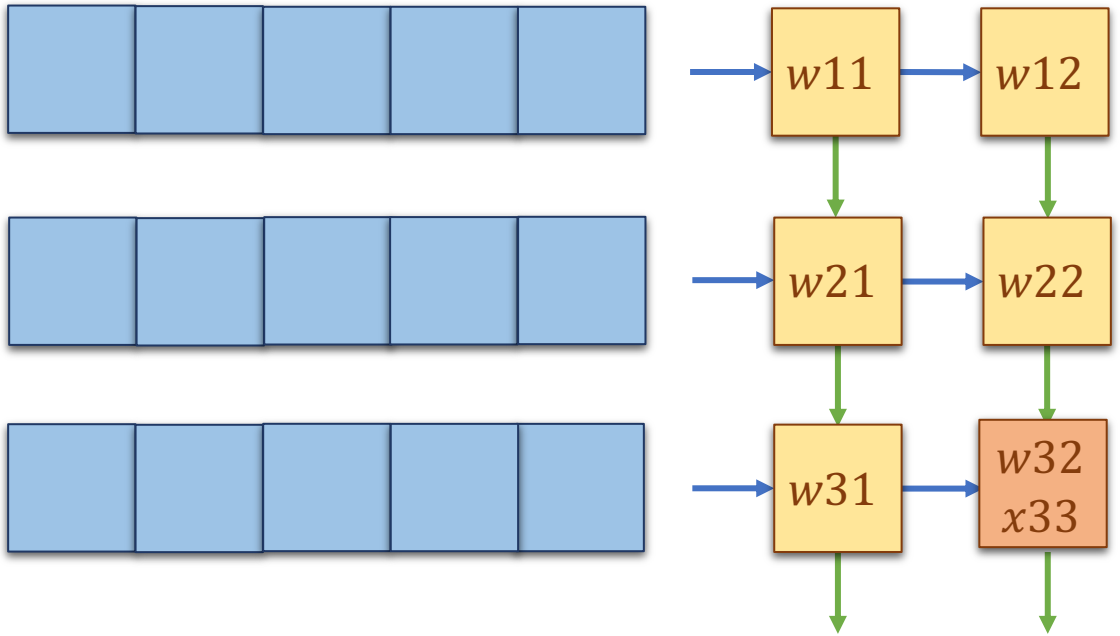
Filtre matrisi

$w11$	$w12$
$w21$	$w22$
$w31$	$w32$

$=$

$$
\begin{aligned}
\boldsymbol{y11} &= w11\,x11 + w21\,x12 + w31\,x13 \\
\boldsymbol{y12} &= w12\,x11 + w22\,x12 + w32\,x13 \\
\boldsymbol{y21} &= w11\,x21 + w21\,x22 + w31\,x23 \\
\boldsymbol{y22} &= w12\,x21 + w22\,x22 + w32\,x23 \\
\boldsymbol{y31} &= w11\,x31 + w21\,x32 + w31\,x33 \\
\boldsymbol{y32} &= w12\,x31 + w22\,x32
\end{aligned}
$$

Sistolik Diziler – Çalışma Prensibi



Aktivasyon matrisi

$x11$	$x12$	$x13$
$x21$	$x22$	$x23$
$x31$	$x32$	$x33$

$\times$

Filtre matrisi

$w11$	$w12$
$w21$	$w22$
$w31$	$w32$

$=$

$$
\begin{aligned}
y11 &= w11\,x11 + w21\,x12 + w31\,x13 \\
y12 &= w12\,x11 + w22\,x12 + w32\,x13 \\
y21 &= w11\,x21 + w21\,x22 + w31\,x23 \\
y22 &= w12\,x21 + w22\,x22 + w32\,x23 \\
y31 &= w11\,x31 + w21\,x32 + w31\,x33 \\
y32 &= w12\,x31 + w22\,x32 + w32\,x33
\end{aligned}
$$

Sistolik Dizilerin Avantaj/Dezavantajları

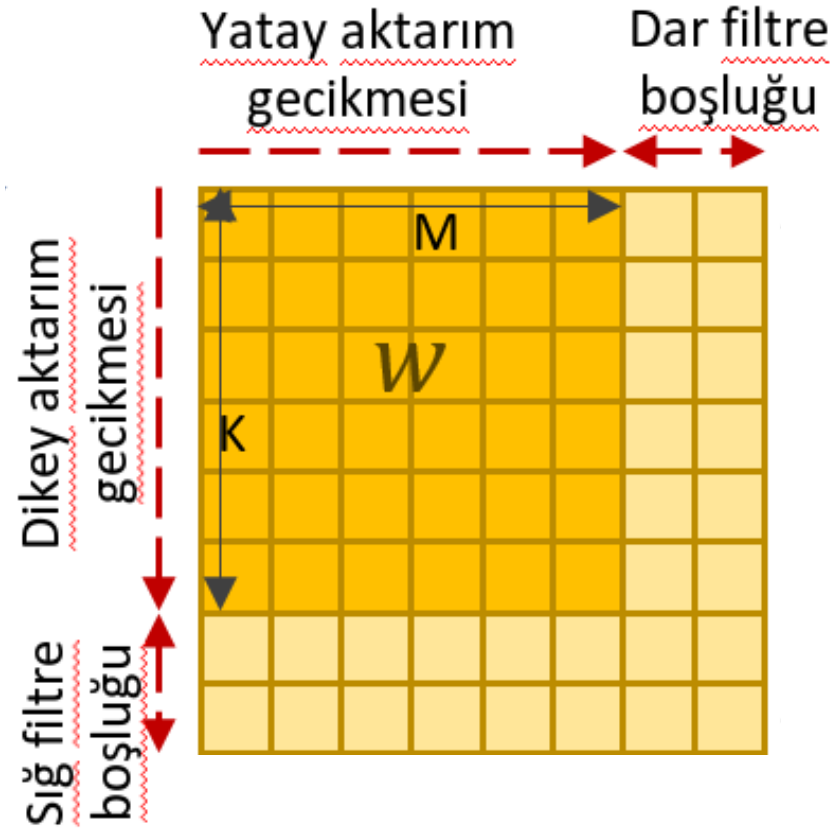
PARALLEL SYSTEMS ARCHITECTURE LAB

✓ Veri aktarımı: komşu işlemci birimleri arasında

- Veri aktarimi için harcanan enerji → 0
- Kısa bağlantı hatları → Ölçeklenebilir tasarım

✗ Dusuk kapasite kullanım oranı

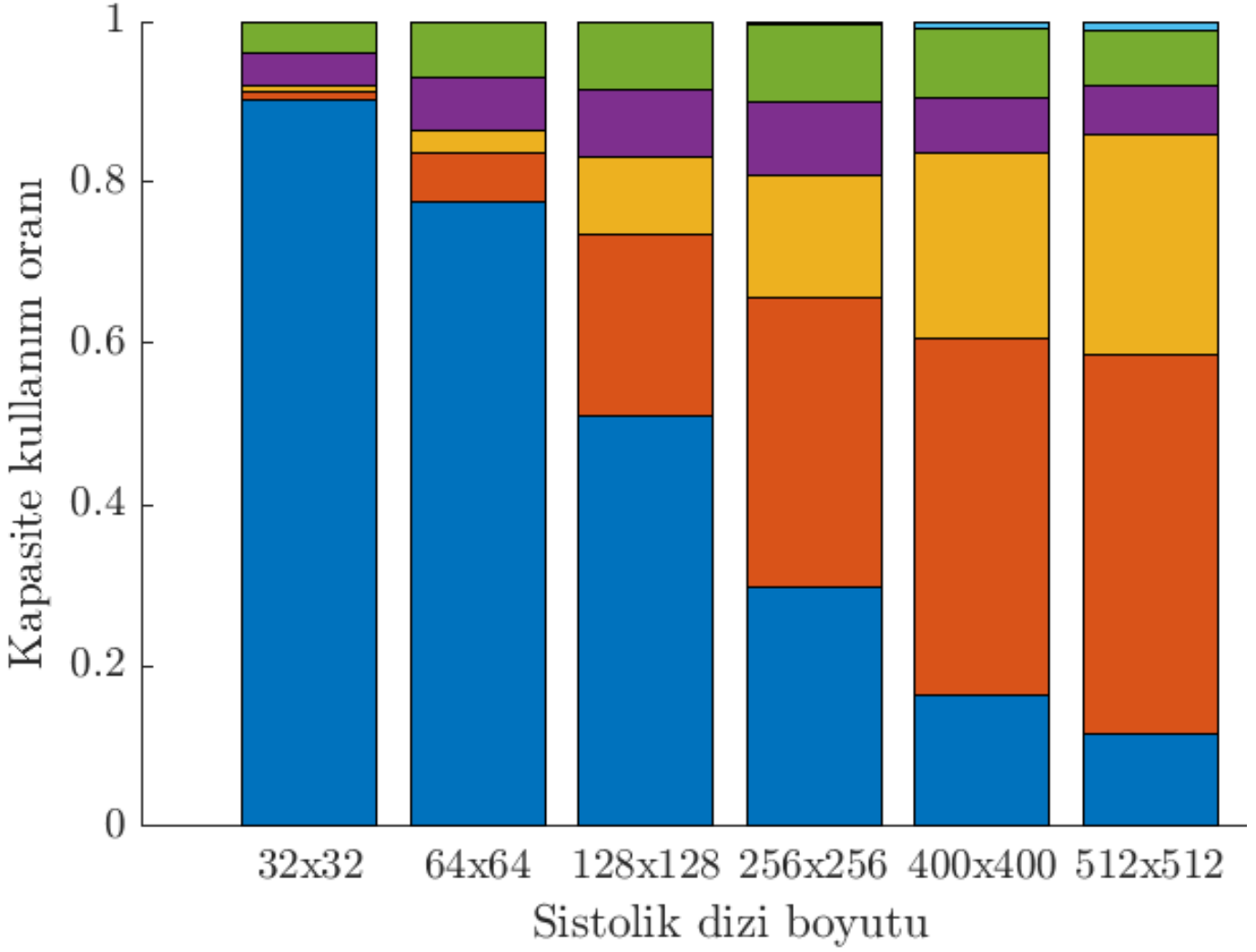
1. Yatay aktarım gecikmesi
2. Dikey aktarım gecikmesi
3. Dar filtre boşluğu
4. Sığ filtre boşluğu



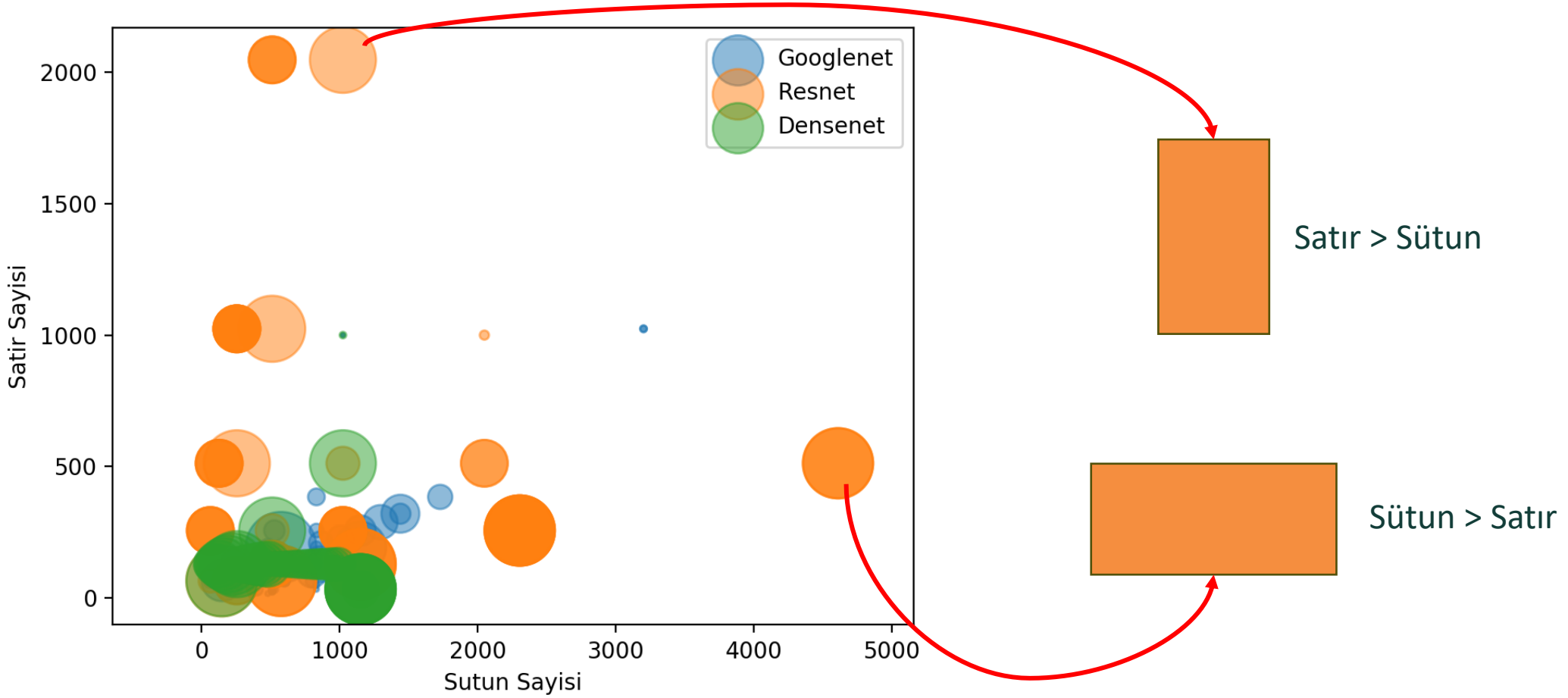
Düşük Kapasite Kullanımı Nedenleri

PARALLEL SYSTEMS ARCHITECTURE LAB

- Büyük sistolik dizi → Düşük kapasite kullanımı
- Kapasitesin >%50'sinin kullanılamama sebebi:
  - Dar/sığ filtre boşluğu



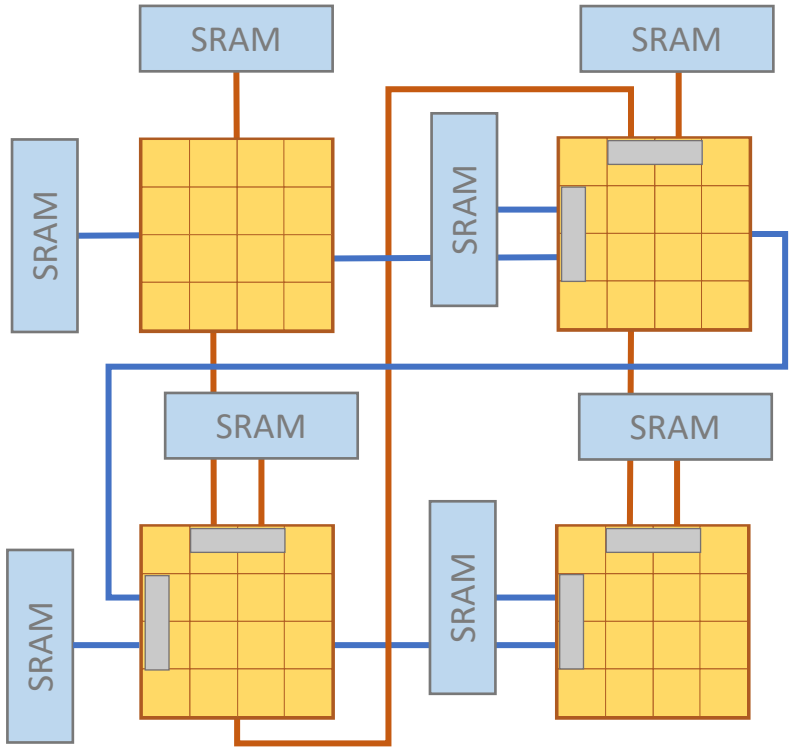
DSA Katman Boyutlarındaki Değişkenlik



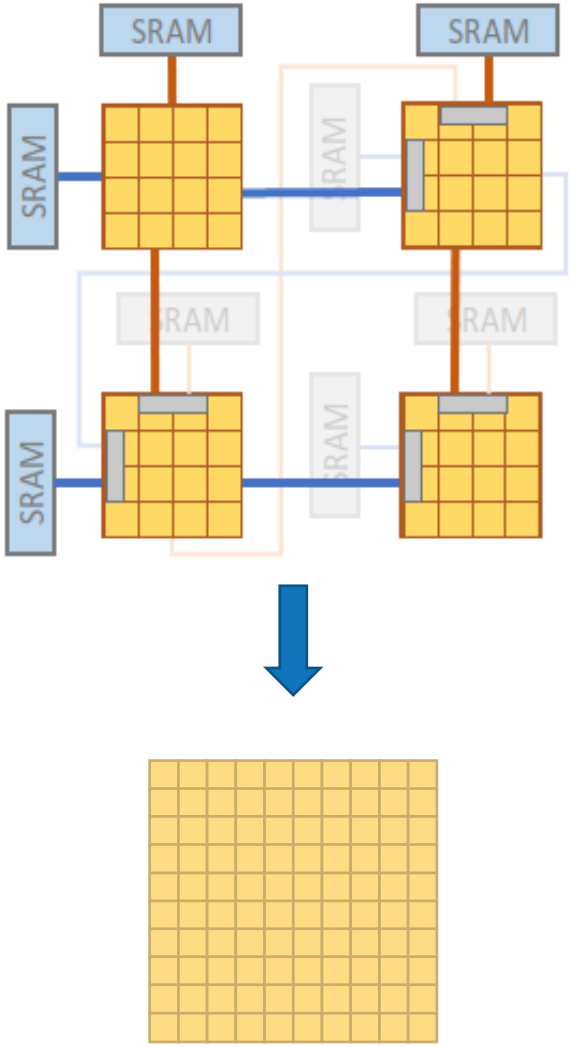
Sabit sistolik dizi + değişken filtre boyutu = düşük kapasite kullanımı

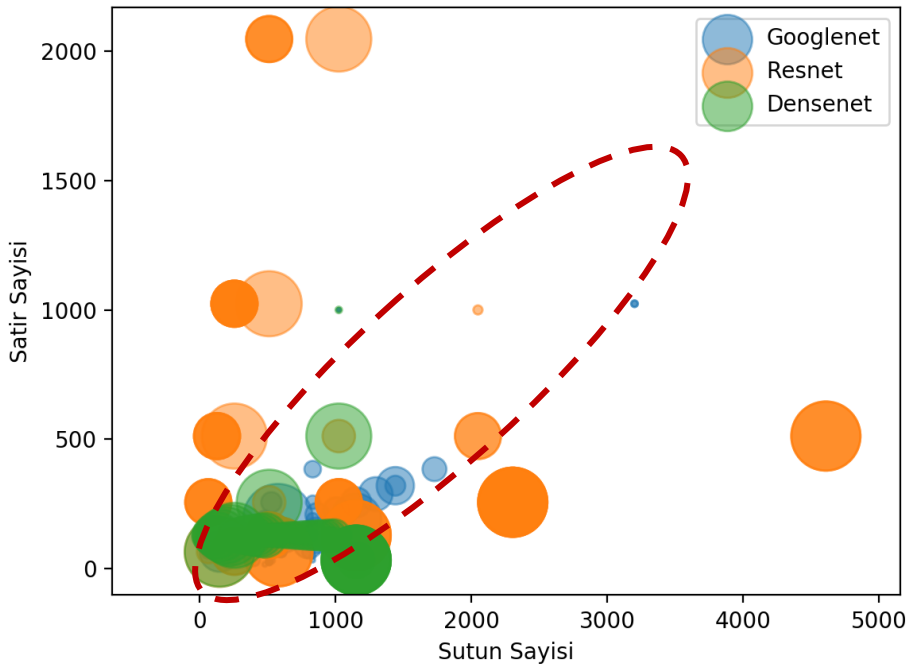
Önerdiğimiz Çözüm

- İri parçalı yeniden yapılandırılabilir mimari (Coarse-grained reconfigurable arch.)
  - Tek büyük parça → Birçok küçük parça
  - Parçalar arası programlanabilir bağlantılar
  - Farklı bağlantı konfigürasyonları → Değişken sistolik dizi boyutu
  - Değişken boyut → Farklı filtrelere daha iyi uyum



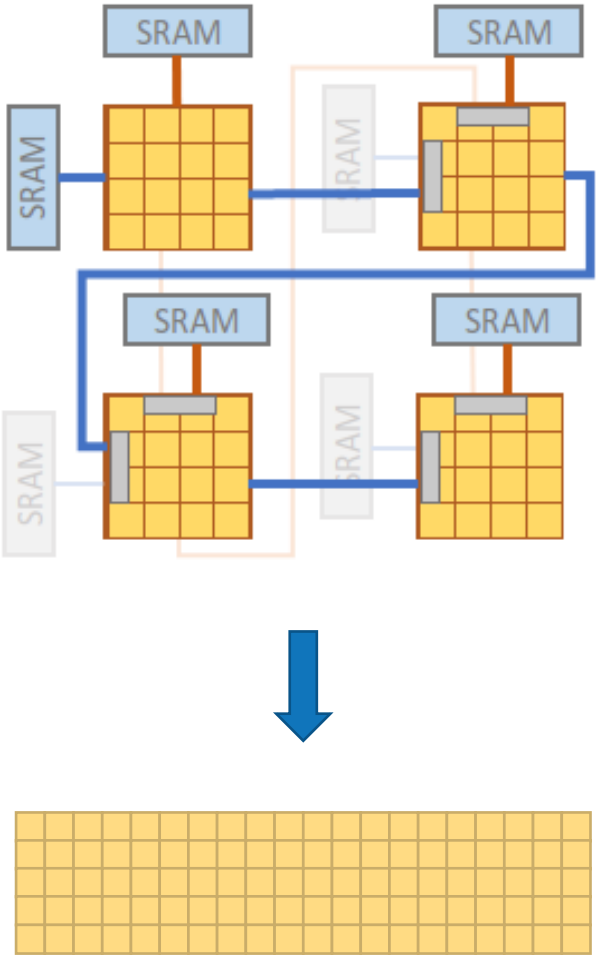
Kare konfigürasyon

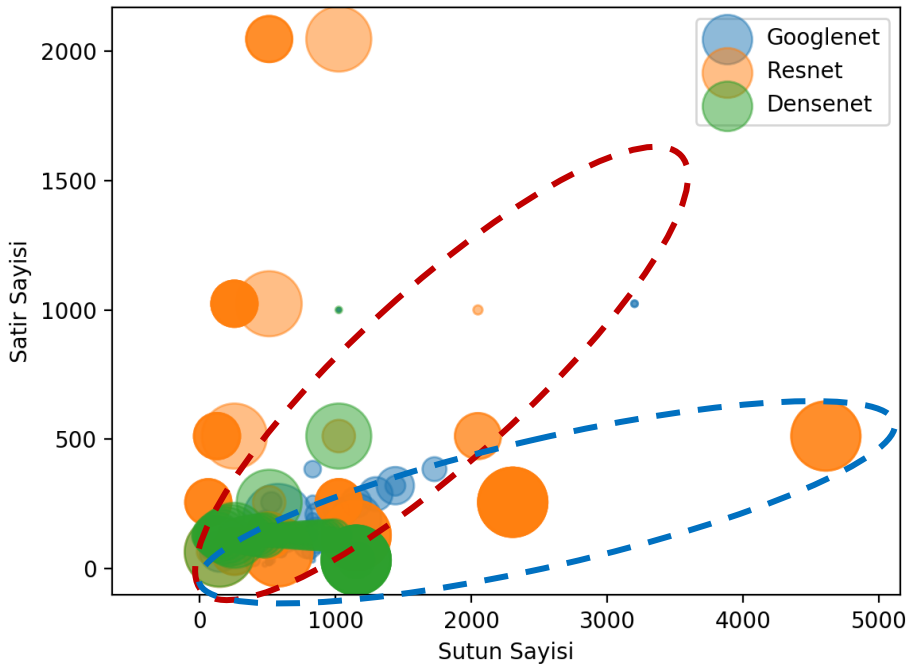




satır ≅ sütun

Geniş konfigürasyon

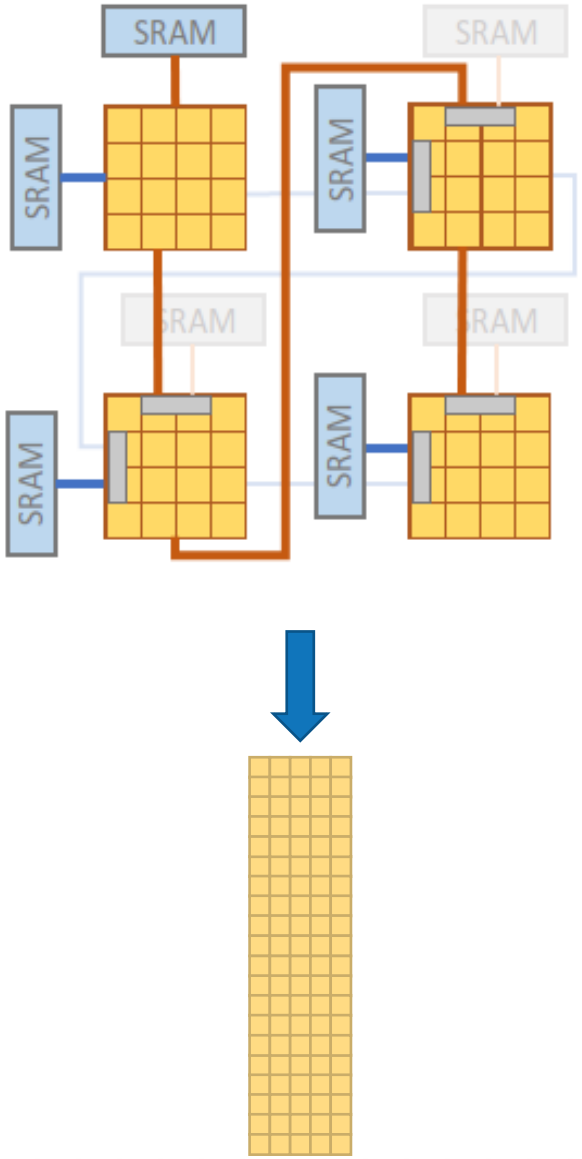


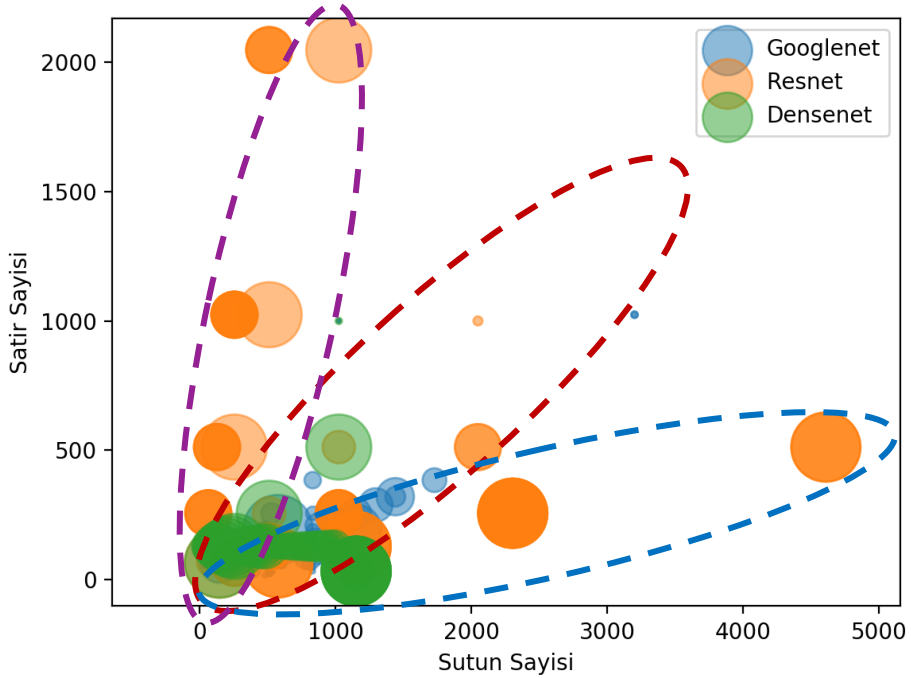


satır ≅ sütun

satır < sütun

Derin konfigürasyon





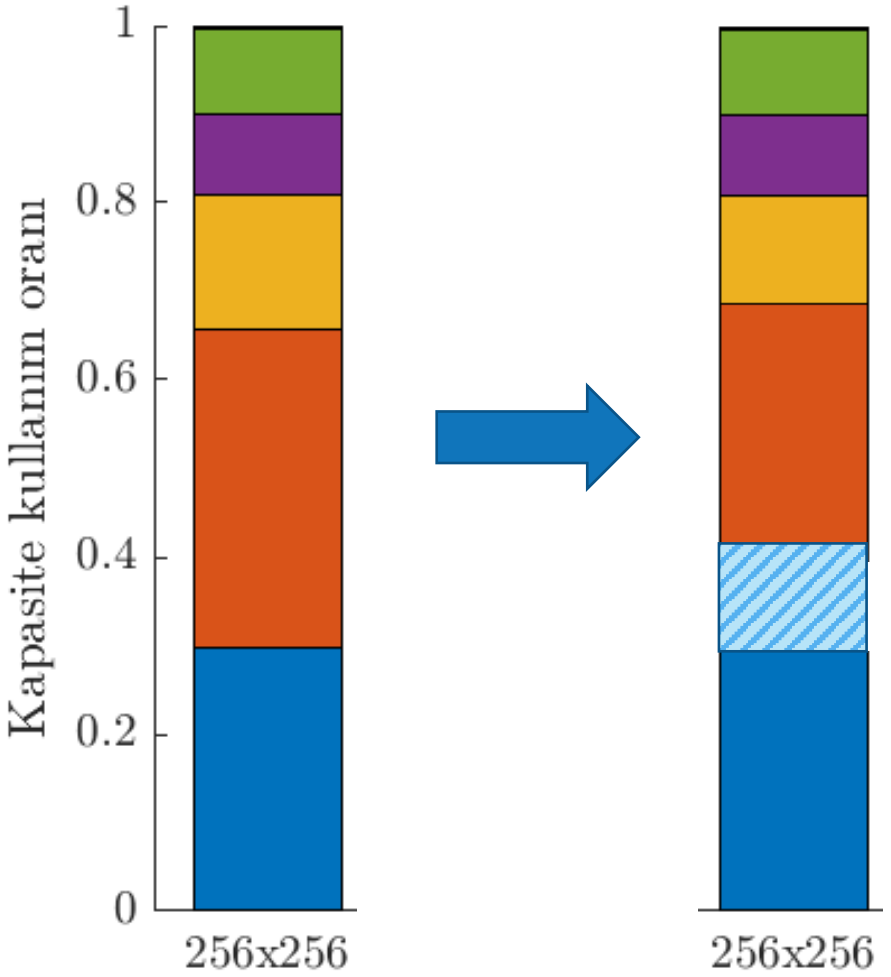
satır ≅ sütun

satır < sütun

satır > sütun

Sonuçlar

- 8 parçaya bölünmüş 256x256'lik bir sistolik dizide
  - Farklı filtre boyutlarına %20 daha iyi uyum
  - %45 daha yüksek etkin işlem gücü



Karşılaştırma

- Google TPU
  - 256x256 standart sistolik dizi
  - 28nm TSMC
  - 36.7 mm^2 silikon alanı
  - %29.7 kapasite kullanım oranı
  - 27.3 TeraOps/s etkin işlem gücü

%6.5 ek yük

%45 artış

- Bu çalışma
  - 8 parçalı 256x256 sistolik dizi
  - 28nm TSMC
  - 39.1 mm^2 silikon alanı
    - İşlemciler arası bağlantı karmaşıklığı
    - Programlanabilirlik icin gerekli çoklayıcılar (multiplexer)
  - %43.3 kapasite kullanım oranı
  - 39.8 TeraOps/s etkin işlem gücü

Kaynakça

- **Sze,2017:** V. Sze et. al., Efficient Processing of Deep Neural Networks: A Tutorial and Survey, Proceedings of the IEEE, Volume: 105 Issue: 12, 2017
- **Suleiman,2017:** A. Suleiman et. al., Towards Closing the Energy Gap Between HOG and CNN Features for Embedded Vision, IEEE International Symposium on Circuits and Systems (ISCAS), 2017
- N. P. Jouppi et. al., In-Datacenter Performance Analysis of a Tensor Processing Unit, International Symposium on Computer Architecture (ISCA), 2017
- H.T. Kung and C. E. Leiserson, "Systolic Arrays (For VLSI)", Sparse Matrix Proceedings, 1978
- C. Szegedy et. al. Going Deeper with Convolutions, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.