# Web Workload characterisation

- Workload: the set of all inputs a system receives over a time period.
- Workload identification → performance evaluation techniques (simulation, benchmarking, etc.)
- Web workload: HTTP message characteristics, Resource characteristics, User behaviour

# Web Workload Parameters

- Protocol: request method, response code
- Resource: content type, resource size, response size, popularity, modification frequency, temporal locality, embedded resources
- user: session inter-arrival times, request inter-arrival times

# HTTP Request methods

- Vast majority of requests: GET method to retrieve documents and execute scripts
- small fraction: use of POST method
- Future trends: WebDAV (Distributed authoring and versioning) may increase PUT/DELETE percentage, tracing/debugging web components may increase TRACE method.

# HTTP response codes

• 200 OK status code: 75% to 90% of responses.

• Next most common response code: 304 Not modified - typically 10% to 30%

• Other popular status codes: redirection (3xx) and client error (4xx).

•Future trends: sophistication level of HTTP clients e.g., 206 partial content

# Web resource characteristics

- Resource size: average resource size for HTML files: 4KB - 8 KB. Image Average Resource Size: 14 KB.
- Wide variability in the HTML file sizes.
- The high variability is captured by the Pareto distribution.

$$F(x) = \left(k / x\right)^{a}, x \geq k$$

- Shape parameter α, and scale parameter k.
- Mean= (kα) / (α-1)
- Pareto distribution is heavy tailed for values of α between 0 and 2.
- Random variables whose distributions are heavy tailed exchibit very high variability: infinite variance, and if $\alpha \leq 1$, infinite mean.

# Web Resource sizes

- Pareto shape parameter, α, for Web resources is between 1.0 and 1.5.
- Pareto is not a good approximation for resources of limited size. Only for large resources.
- Body of distribution is better modeled through the lognormal distribution.

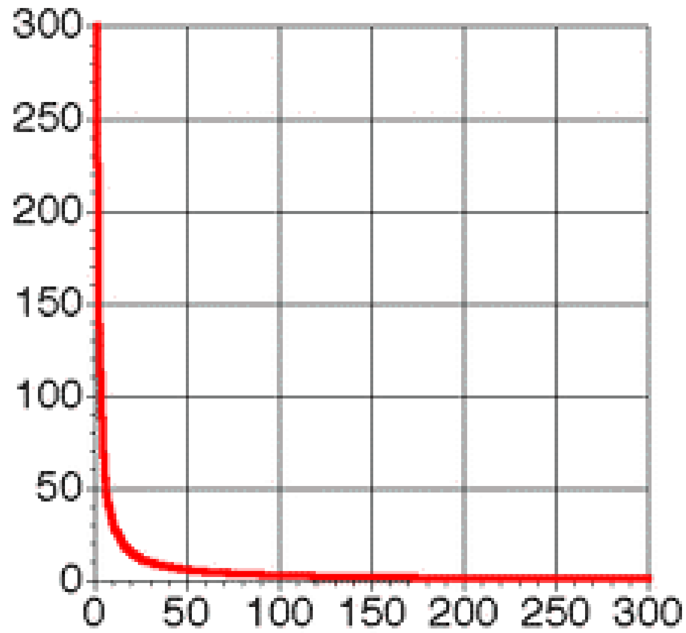
# Zipf Law

Zipf's law usually refers to the 'size' ***y*** of an occurrence of an event relative to it's rank ***r***. George Kingsley Zipf, a Harvard linguistics professor, sought to determine the 'size' (or frequency of use in English text) of the 3rd or 8th or 100th most common word. **Zipf's law states that the size of the r'th largest occurrence of the event is inversely proportional to it's rank**:

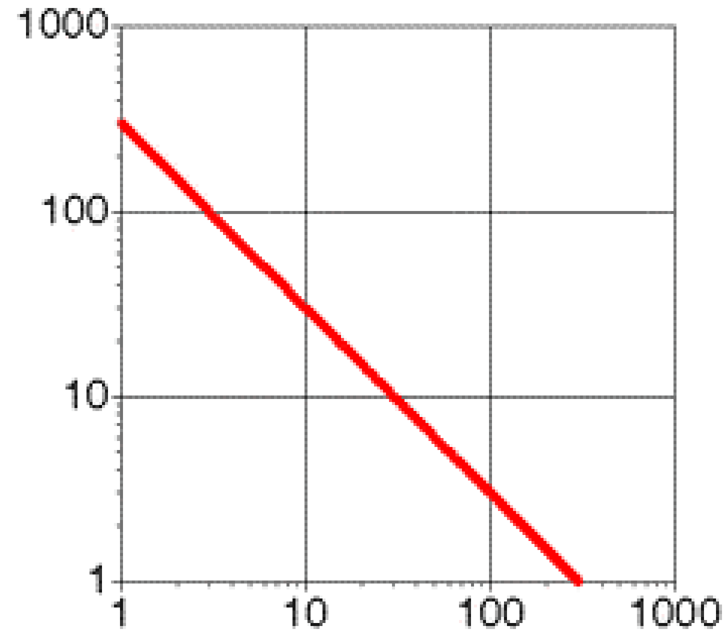
$y \sim r^{-b}$, with ***b*** close to unity.

**GK Zipf**, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, 1949).

# Zipf Distribution



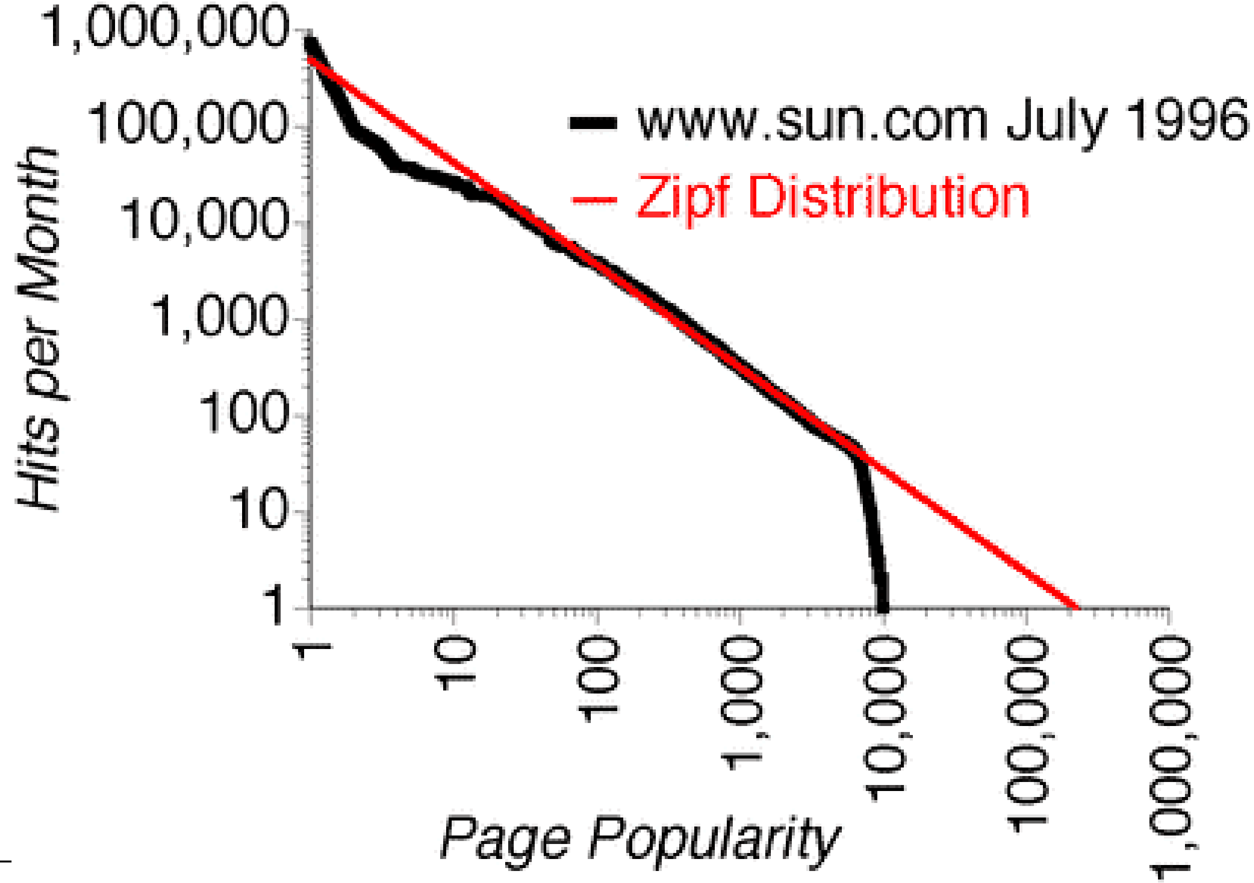

*Linear scales on both axes*




*Logarithmic scales on both axes*

The same data plotted on linear and logarithmic scales. Both plots show a Zipf distribution with 300 datapoints.

# Zipf Distribution



# Pareto Law

Pareto was interested in the distribution of income. Instead of asking what the ***r*** th largest income is, he asked how many people have an income greater than ***x***. Pareto's law is given in terms of the cumulative distribution function (CDF), i.e. the number of events larger than ***x*** is an inverse power of ***x***:

$$P[X > x] \sim x^{-k}.$$

It states that there are a few multi-billionaires, but most people make only a modest income.

# Pareto Distribution

The Pareto distribution gives the probability that a person's income is greater than or equal to x:

$Pr[X >= x] = (m/x)^k, m > 0, k > 0, x >= m$,

where ***m*** represents a minimum income. As a consequence, the CDF

$Pr[X < x] = 1 - (m/x)^k$

and the PDF is

$p_X(x) = k\, m^k x^{-(k+1)}, m > 0, k > 0, x >= m$

# Temporal locality

- The time between successive requests for the same resource.
- Temporal locality captures the likelihood that a requested resource will be requested again in the near future.
- Measurement: stack distance.
- (a,b,a,b,a,b,a) vs (a,a,a,a,b,b,b): first stream exhibits lower temporal locality than the second.
- Small stack distances: higher temporal locality.
- Distribution of temporal locality: lognormal distribution.

# Embedded References

- Web pages have a median of 8 to 20 embedded references.
- The respective distribution has high variability, following the pareto distribution.

# Session inter-arrival times.

- Session inter-arrival times follow the exponential distribution.
- Exponential model is not adequate for interarrival times of TCP connections and HTTP requests.

# Web traffic models

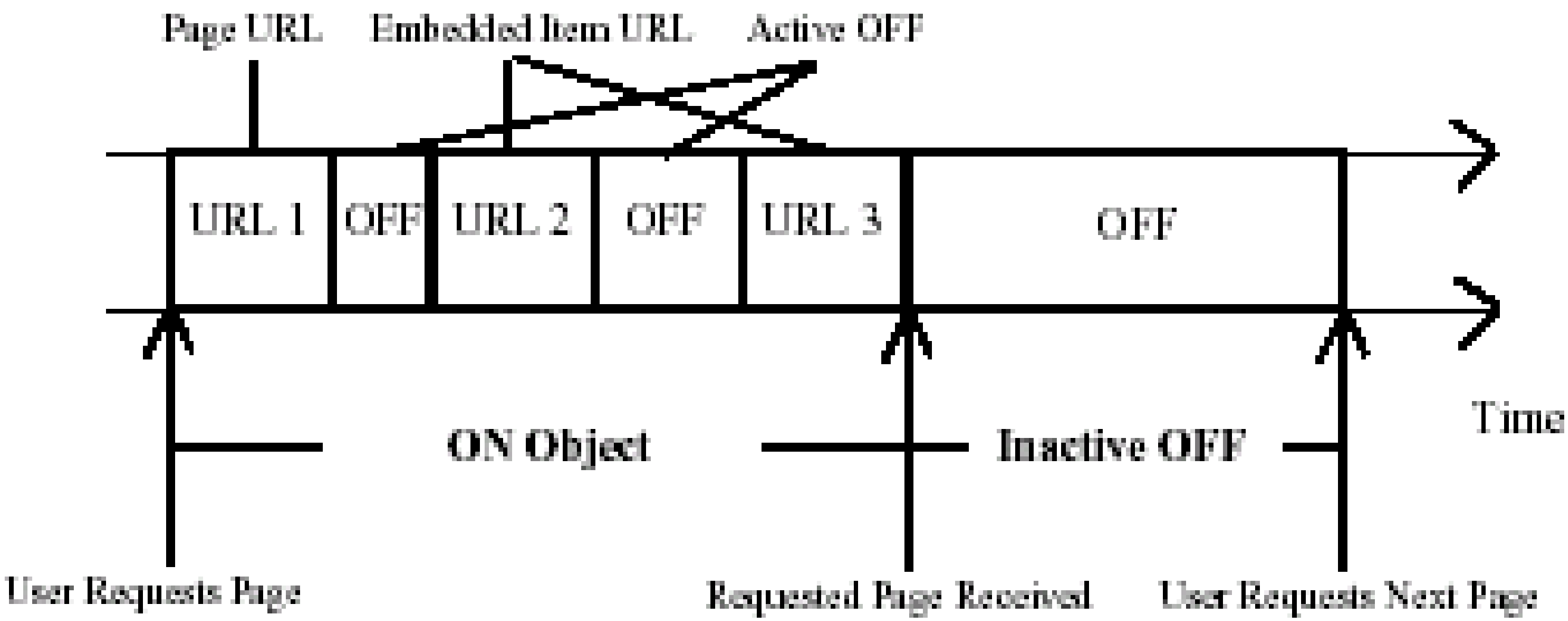
• WWW traffic exhibits characteristics consistent with self-similar traffic models.

• Self-similarity is attributed to the multiplexing of a large number of ON/OFF sources where both the ON and the OFF period lengths are heavy tailed processes.

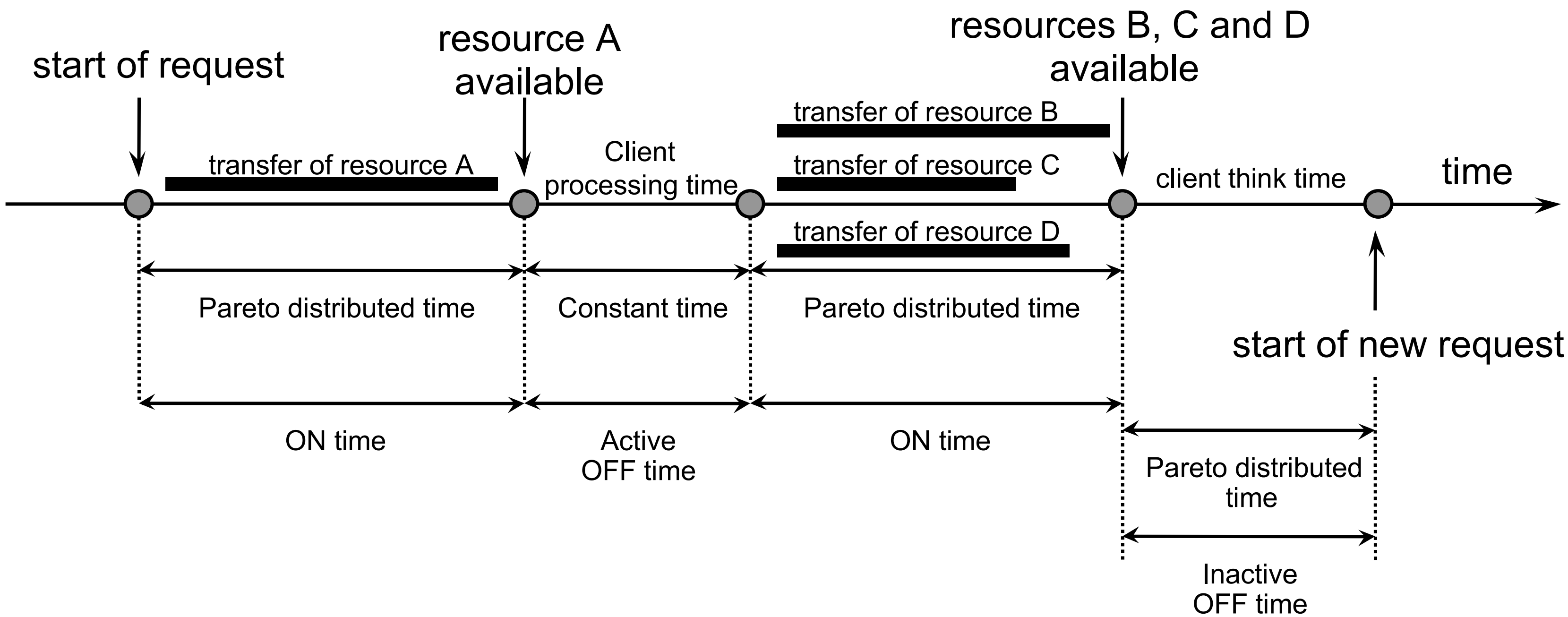
• ON times correspond to WWW resource transmissions while the OFF times correspond to intervals of client/browser inactivity.

• OFF times are classified either as Active (attributed to client processing delays: parsing, rendering) or as Inactive (user think time).

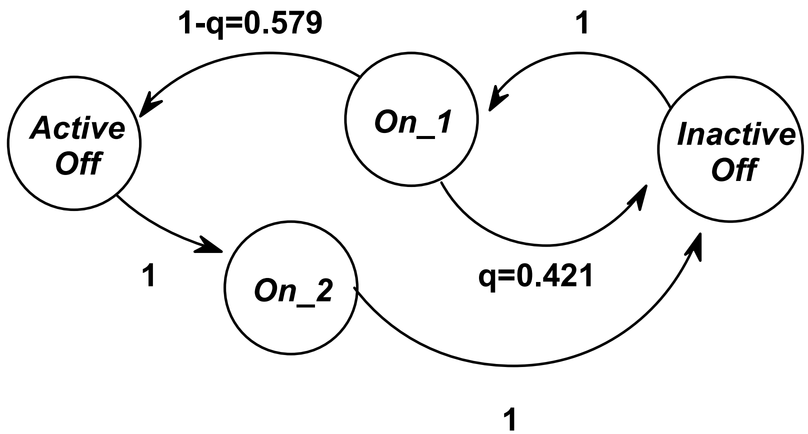
# Web traffic model



# Web traffic model



# Web traffic chain




- On_1: HTML page retrieval
- On_2: retrieval of embedded references
- Inactive Off ~ Pareto
- On_1 ~ Pareto
- On_2 ~ Pareto
- Active Off ~ Weibull

# Web model statistics

| Component | Model | Probability Density Function | Parameters |
|---|---|---|---|
| File Sizes – Body | Lognormal | $p(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x - \mu)^2/2\sigma^2}$ | $\mu = 9.357$; $\sigma = 1.318$ |
| File Sizes – Tail | Pareto | $p(x) = \alpha k^{\alpha} x^{-(\alpha+1)}$ | $k = 133K$; $\alpha = 1.1$ |
| Popularity | Zipf | | |
| Temporal Locality | Lognormal | $p(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x - \mu)^2/2\sigma^2}$ | $\mu = 1.5$; $\sigma = 0.80$ |
| Request Sizes | Pareto | $p(x) = \alpha k^{\alpha} x^{-(\alpha+1)}$ | $k = 1000$; $\alpha = 1.0$ |
| Active OFF Times | Weibull | $p(x) = \frac{b x^{b-1}}{a^b} e^{-(x/a)^b}$ | $a = 1.46$; $b = 0.382$ |
| Inactive OFF Times | Pareto | $p(x) = \alpha k^{\alpha} x^{-(\alpha+1)}$ | $k = 1$; $\alpha = 1.5$ |
| Embedded References | Pareto | $p(x) = \alpha k^{\alpha} x^{-(\alpha+1)}$ | $k = 1$; $\alpha = 2.43$ |

# Self-Similar processes

Traffic processes are said to be self-similar if they look qualitatively the same irrespective of the time scale from which we look at them.