yoast

Crawler killer: migliorare il mondo un robot alla volta

Paolo L. Scala – WordCamp Torino 2023 – 11, 12 Aprile

Outline

- Contesto
- Il problema
- Strategie di mitigazione
- Un’architettura alternativa
- Key takeaways

Contesto

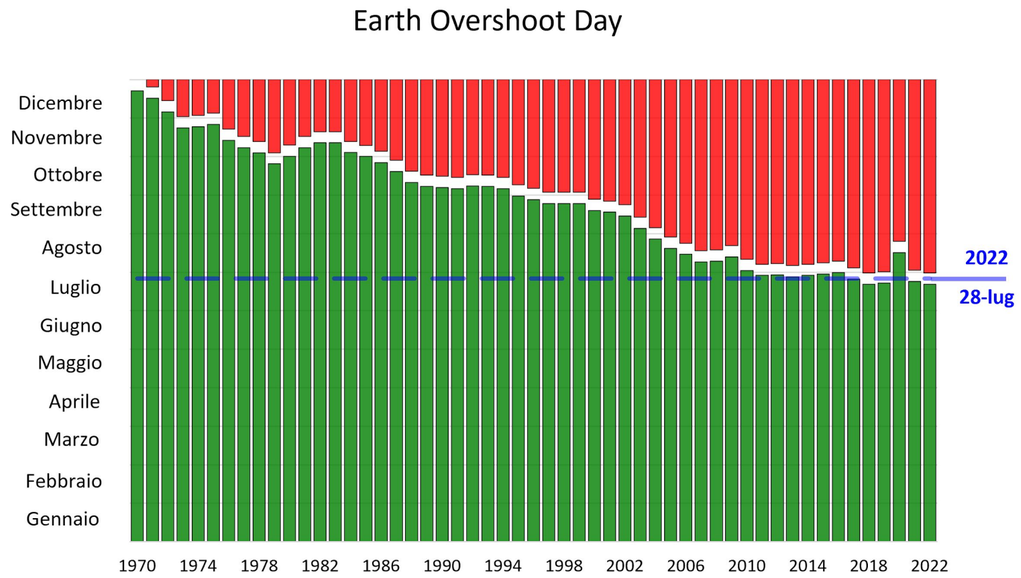
Contesto

- Impatto sull’ambiente del WWW
- Prospettiva server Web
- Energia elettrica in rapporto alle richieste

Il problema

Il problema

- Earth Overshoot day 2022: 28 Luglio



Il problema

- Carbon footprint
- *the life cycle carbon equivalent emissions and effects related to a product or service*
- Nel 2016, il World Economic Forum ha classificato il riscaldamento globale come la minaccia numero 1 per la società e l'economia.

(Malmodin, J.; Lundén, D. The Energy and Carbon Footprint of the Global ICT and E&M Sectors 2010–2015. Sustainability 2018, 10, 3027. https://doi.org/10.3390/su10093027)

Il problema

~ 2.3TW → ~ 20150 TWh per anno

(https://www.carbonfootprint.com)

L’mpatto del WWW

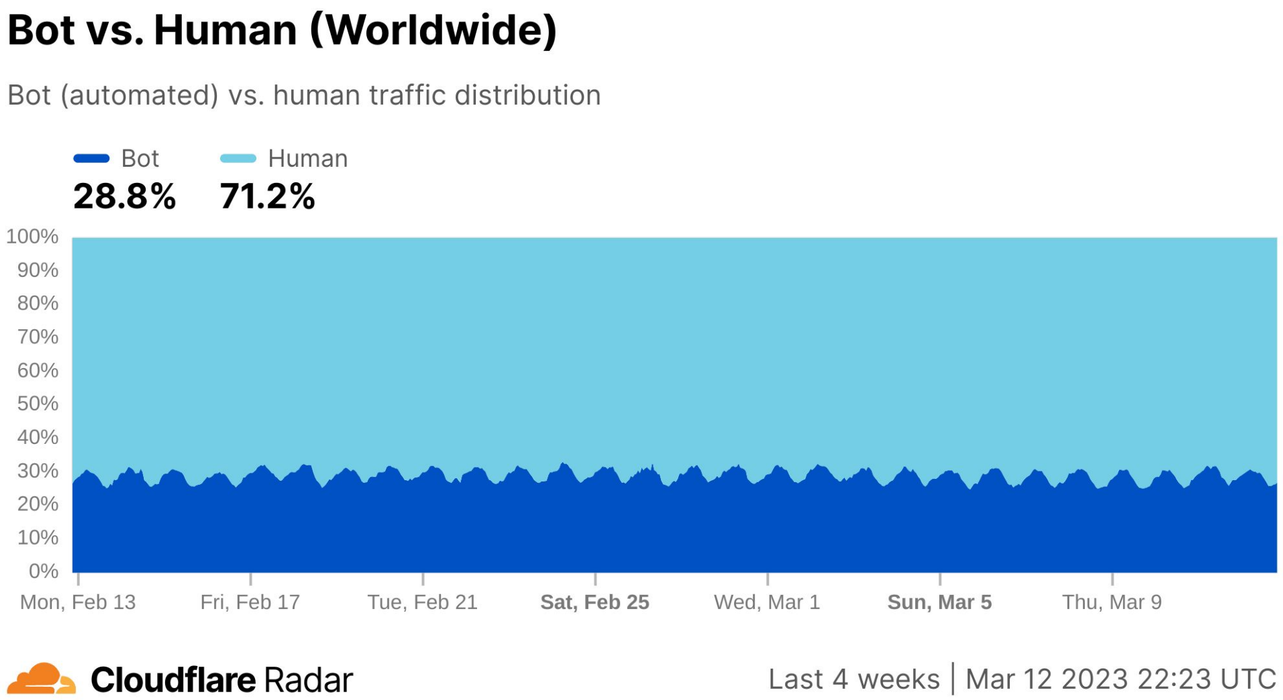
- Qual è il contributo della Rete?
- ~3.6% - 6.2% → 84-143GW → ~736 – 1250GWh in un anno

Category	Wall-socket power	Wall-socket duty cycle	Total power (min)		Total power (max)	
			Wall-socket	Embodied	Wall-socket	Embodied
Desktops	150 W	0.5	28.1 GW	22.3 GW	53.4 GW	42.3 GW
Laptops	40 W	0.5	11.3 GW	26.7 GW	15.0 GW	35.6 GW
Cloud	450 W	1.0	18.0 GW	2.1 GW	22.5 GW	2.6 GW
Smartphones	1 W	0.5	0.13 GW	4.0 GW	0.45 GW	14.3 GW
Servers	375 W	1.0	18.8 GW	2.6 GW	35.6 GW	5.0 GW
Routers	5 kW	1.0	4.5 GW	0.48 GW	5.0 GW	0.53 GW
Wi-Fi/LAN	20 W	1.0	1.5 GW	0.80 GW	2.0 GW	1.1 GW
Cell Towers	3 kW	1.0	1.5 GW	0.16 GW	7.5 GW	0.80 GW
Telecom Switches	75 kW	1.0	0 GW	0 GW	1.4 GW	0.06 GW
Fiber Optics	0 W	0	0 GW	23.8 GW	0 GW	42.8 GW
Copper	0 W	0	0 GW	3.7 GW	0 GW	18.5 GW
Total for Internet			84 GW	87 GW	143 GW	164 GW
			170 GW		**307 GW**	

(Barath Raghavan and Justin Ma. 2011. The energy and emergy of the internet. In Proceedings of the 10th ACM Workshop on hot topics in networks. 1–6.)

L’impatto dei server Web

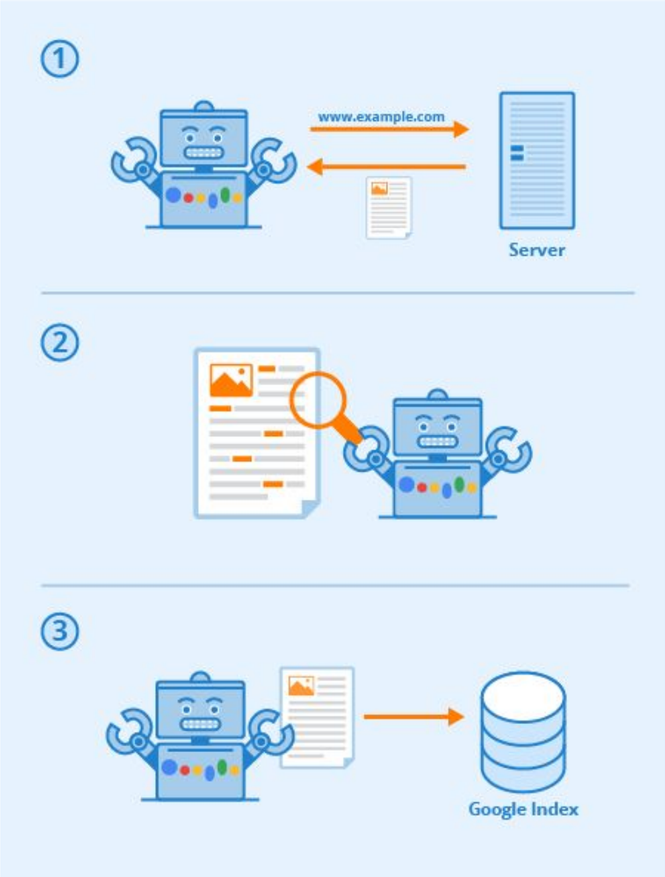
- L’energia consumata da un server Web è proporzionale alle richieste che riceve



Cos’è un bot?

- ***Crawler*** *(sometimes also called a* ***robot*** *or* ***spider****) is a generic term for any* ***program that is used to automatically discover and scan websites*** *by following links from one web page to another*
- Agente alla base del funzionamento di ogni motore di ricerca

Come funziona un bot?



Come funziona un bot?

- Fase 0: URL Discovery
- Non esiste un «registro globale delle pagine Web»
- Scoperte attraverso link di pagine che Google già conosce
- Scoperte attraverso una sitemap fornita direttamente a Google

- Fase 1: Crawling
- Googlebot utilizza un algoritmo per capire di quali siti fare crawling, quanto spesso e quante pagine deve visitare

Come funziona un bot?

- Fase 2: Rendering
- La pagina viene rendererizzata, JavaScript compreso
- Il contenuto viene analizzato

- Fase 3: Indexing
- La pagina viene inserita nel Google Index
- Metriche utili al ranking
- Metadati

Come funziona un bot?

- Il crawling di un URL da parte del GoogleBot continua per sempre
- Motivo: aggiornare il Google Index

- GoogleBot ri-visita un URL una volta ogni 4-30 giorni
- Dipende quanto il sito è attivo

La Domanda

E’ possibile limitare questo traffico?

Strategie di mitigazione

I metodi «tradizionali»

- robots.txt
- File che istruisce i crawler su quali URL possono visitare e quali no
- robots meta tag
- Meta tag specificabile nell'elemento html <head> di una pagina a cui è associabile (tra gli altri) il valore noindex/nofollow

I metodi «tradizionali»

- Google Search Console
- È possibile limitare la «crawl rate»

- Attributo rel
- Attributo opzionale dei tag html <a> a cui può essere assegnato il valore nofollow

I metodi «tradizionali»: esempi

- robots.txt

User-agent: Googlebot

Disallow: /nogooglebot/

User-agent: *

Allow: /

Sitemap: https://www.example.com/sitemap.xml

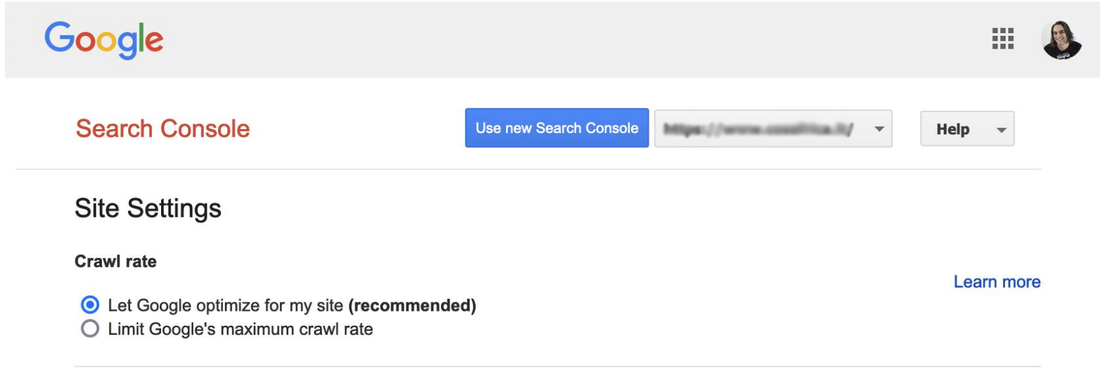
I metodi «tradizionali»: esempi

- robots meta tag

```
<head>
    <meta charset="UTF-8" />
    <meta name='robots' content='index, follow,
        max-image-preview:large,
        max-snippet:-1,
        max-video-preview:-1' />
</head>
```

I metodi «tradizionali»: esempi

- Google Search Console



I metodi «tradizionali»: esempi

- Attributo **rel**

```
<a rel=‘nofollow’ href=’https://www.test.com’>
    Example link
</a>
```

I metodi «tradizionali»: Limitazioni

- robots.txt
- Non sempre è possibile modificarlo direttamente
- Non alla portata di tutti

- robots metatag
- Da specificare pagina per pagina
- Limitato a risorse HTML
- Non alla portata di tutti

I metodi «tradizionali»: Limitazioni

- Google Search Console
- Assenza di granularità nel controllo

- Attributo rel
- Efficacia limitata
- Da specificare per ogni URL

Un approccio diverso

- Specifico per WordPress
- Prevede l’utilizzo di *hook* per evitare la creazione di URL spesso inutilizzati

URL inutilizzati: feed RSS

- Cos’è RSS

- Really Simple Syndication
- standard per la pubblicazione e distribuzione di contenuti aggiornati
- Un RSS feed è uno stream XML pubblicato da un content provider
- Chi è interessato a ricevere aggiornamenti relativi alla pubblicazione di contenuti si abbona al feed
- Utili per news aggregation

URL inutilizzati: feed RSS

- Global feed
- https://basic.wordpress.test/feed/

- Global comment feed
- https://basic.wordpress.test/comments/feed

- Post comment feed
- https://basic.wordpress.test/post-example/feed/

URL inutilizzati: feed RSS

- Author feed
- https://basic.wordpress.test/author/paolo/feed/

- Post type feed
- htts://basic.wordpress.test/post-type-example/feed/

- Category feed
- htts://basic.wordpress.test/category/category-example/feed/

URL inutilizzati: feed RSS

- Tag feed
- https://basic.wordpress.test/tag/tag-example/feed/

- Custom taxonomy feed
- https://basic.wordpress.test/custom/taxonomy/feed/

- Search result feed
- https://basic.wordpress.test/search/world/feed/rss2/

URL inutilizzati: feed RSS

- **Feed Atom/RDF**

- http://basic.wordpress.test/feed/atom
- http://basic.wordpress.test/feed/rdf
- http://basic.wordpress.test/comments/feed/atom
- http://basic.wordpress.test/comments/feed/rdf
- http://basic.wordpress.test/hello-world/feed/atom
- http://basic.wordpress.test/hello-world/feed/rdf

URL inutilizzati: oEmbed

- Standard per l’embedding di risorse Web
- API molto semplice
- Permette di fruire una risorsa senza visitarla direttamente
- Esempio di richiesta:

http://www.flickr.com/services/oembed/?format=json&url=http%3A//www.flickr.com/photos/bees/2341623661

- Esempio di risposta:


```
{       "version": "1.0",
        "type": "photo",
        "width": 240,
        "height": 160,
        "title": "ZB8T0193",
        "url": "http://farm4.static.flickr.com/3123/2341623661_7c99f48bbf_m.jpg",
        "author_name": "Bees",
        "author_url": "http://www.flickr.com/photos/bees/",
        "provider_name": "Flickr",
        "provider_url": http://www.flickr.com/
}
```

URL inutilizzati: oEmbed

- Nel caso di un post WordPress

- <link rel="alternate" type="application/json+oembed" href="http://basic.wordpress.test/wp-json/oembed/1.0/embed?url=http%3A%2F%2Fbasic.wordpress.test%2F2022%2F05%2Fhello-world%2F" />

- <link rel="alternate" type="text/xml+oembed" href="http://basic.wordpress.test/wp-json/oembed/1.0/embed?url=http%3A%2F%2Fbasic.wordpress.test%2F2022%2F05%2Fhello-world%2F&format=xml" />

- Stesso contenuto, cambia il formato della risposta (JSON vs XML)

URL inutilizzati: JSON Rest

- Interfaccia che permette ad applicazioni terze di interagire con WordPress
- L’interazione avviene attraverso lo scambio di messaggi JSON
- WordPress pubblica degli URL, detti **endpoint**, che rappresentano le varie risorse con cui è possibile interagire (post, tassonomie, pagine, ecc.)
- <link rel="https://api.w.org/" href="http://basic.wordpress.test/wp-json/" />

URL inutilizzati: shortlink

- Versione accorciata di un URL
- <link rel='shortlink' href='http://basic.wordpress.test/?p=1' />

Una soluzione «manuale»

- Utilizzando gli hook di WordPress e la direttiva remove_action

- Feed RSS

  remove_action('wp_head', 'feed_links_extra', 3);

- oEmbed

  remove_action('wp_head', 'wp_oembed_add_discovery_links');

Una soluzione «manuale»

- REST API

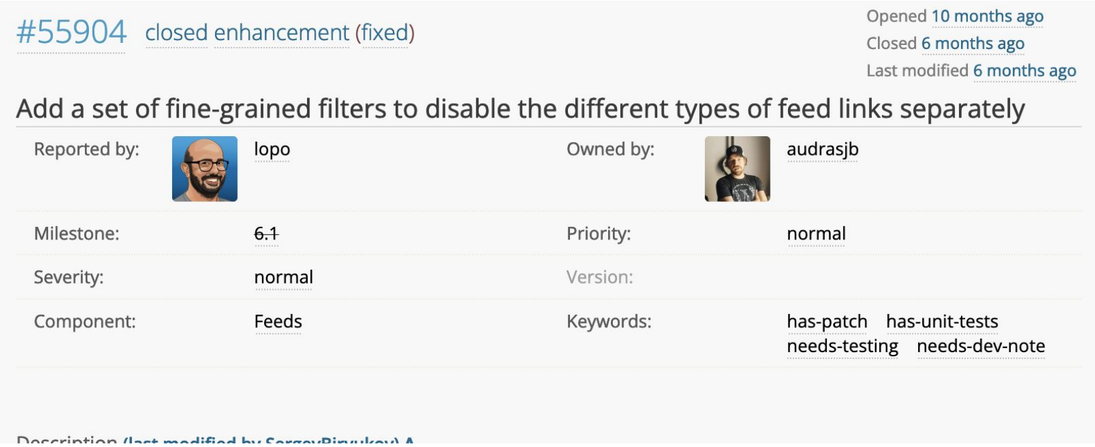
```
remove_action('wp_head','rest_output_link_wp_head' );
remove_action('template_redirect', 'rest_output_link_header', 11 )
```

- Shortlink

```
remove_action( 'wp_head', 'wp_shortlink_wp_head' );
remove_action('template_redirect','wp_shortlink_header', 11 );
```

Una soluzione «manuale»

- Feed RSS: proposta per un controllo più granulare
- Patch presentata durante il WCEU 2022 da Enrico Battocchi (aka Lopo)



#55904 closed enhancement (fixed)

Opened 10 months ago
Closed 6 months ago
Last modified 6 months ago

Add a set of fine-grained filters to disable the different types of feed links separately

Reported by:	lopo	Owned by:	audrasjb
Milestone:	~~6.1~~	Priority:	normal
Severity:	normal	Version:	
Component:	Feeds	Keywords:	has-patch has-unit-tests needs-testing needs-dev-note

- https://core.trac.wordpress.org/ticket/55904
- Disponibile da WordPress 6.1

Una soluzione «manuale»

- Aggiunge i seguenti hook:
- feed_links_extra_show_post_comments_feed
- feed_links_extra_show_post_type_archive_feed
- feed_links_extra_show_category_feed
- feed_links_extra_show_tag_feed
- feed_links_extra_show_tax_feed
- feed_links_extra_show_author_feed
- feed_links_extra_show_search_feed

Soluzione automatica: plugin

- oEmbed
- https://wordpress.org/plugins/disable-embeds/

- REST API
- https://wordpress.org/plugins/disable-wp-rest-api/

- All-in-one
- https://perfmatters.io/
- https://yoast.com/

Un’architettura alternativa

Un’architettura alternativa

- Attualmente
- I motori di ricerca visitano gli URL indipendentemente dai loro aggiornamenti (Pull)

- Approccio contrario
- I content provider segnalano ai motori di ricerca quando un URL presenta del contenuto aggiornato (Push)

- Vantaggi
- Maggior visibilità
- Minor traffico inutile

Un’architettura alternativa: IndexNow

- Protocollo che consente di inviare ai motori di ricerca un ping relativo ad un URL che è stato cambiato
  - Contenuto modificato
  - Contenuto creato
  - Contenuto cancellato
- API molto sempice
  - https://<searchengine>/indexnow?url=url-changed&key=your-key

IndexNow: vantaggi

- Supportato da più motori di ricerca
  - Yandex
  - Bing
  - Seznam.cz
  - Google sta valutando se supportare o meno il protocollo
- Permette di utilizzare un solo ping per più URL che sono cambiati
- Il ping si propaga a tutti i motori di ricerca che suppportano IndexNow

Key takeaways

Key take-aways

- Una porzione non banale di traffico Web è dovuta ai robot
- Questa attività ha un impatto non banale sul consumo energetico
- E quindi sull’ambiente

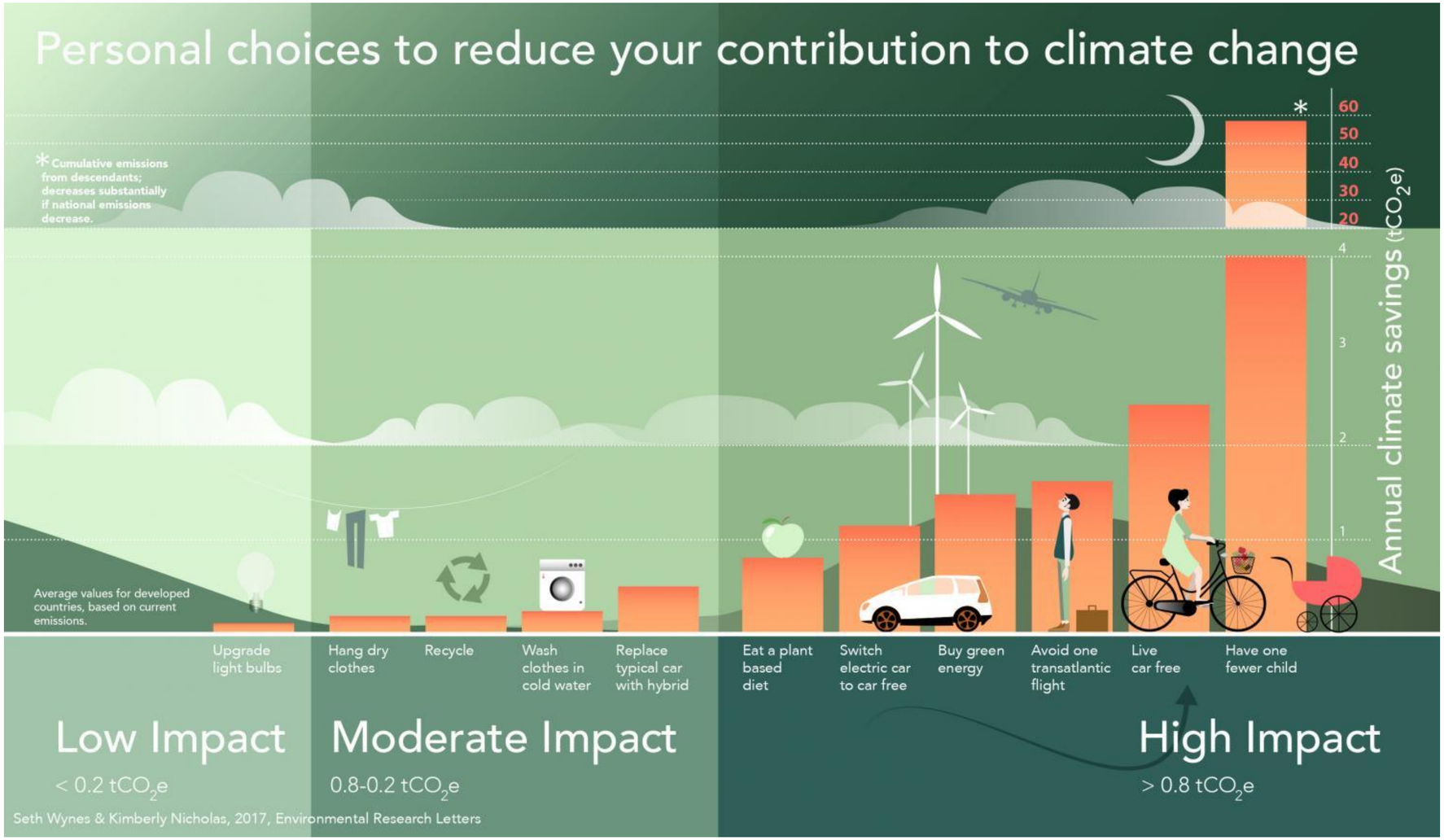
Key take-aways

- Possiamo limitare tale traffico
- Aumento di performance
- Minore energia consumata
- Prospettive future positive

Key take-aways

Cambieremo davvero il mondo?

Key take-aways



Credit: Seth Wynes/Kimberly Nicholas, *Environmental Research Letters*, 2017

Key take-aways

”Un viaggio di mille miglia inizia con un singolo passo”

(S. Berlusconi)

Credit: Seth Wynes/Kimberly Nicholas, *Environmental Research Letters*, 2017

Key take-aways

”Un viaggio di mille miglia inizia con un singolo passo”

~~(S. Berlusconi)~~
(Lao Tzu)

Credit: Seth Wynes/Kimberly Nicholas, *Environmental Research Letters*, 2017


YOASTWAY